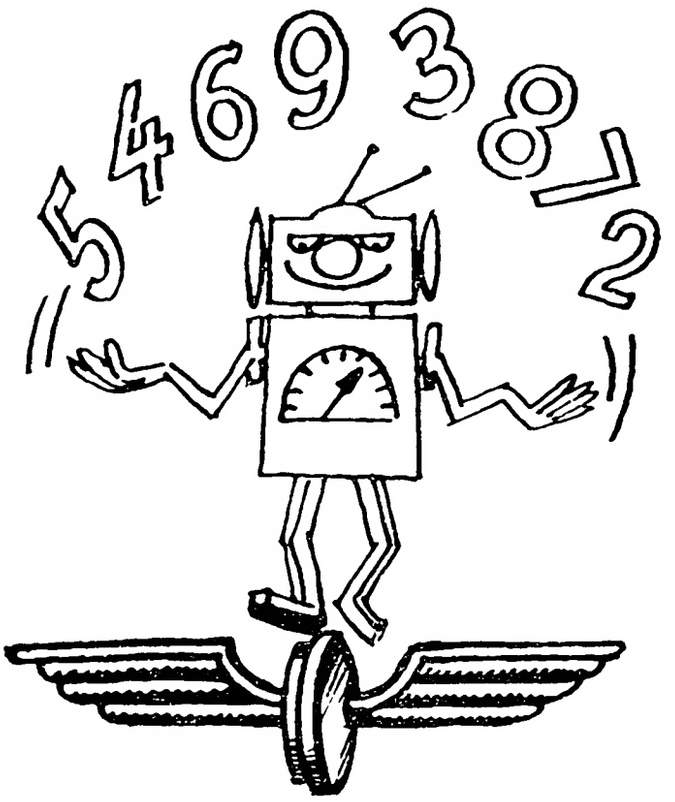
Г.А.Платонов
М.А.Файнберг
М.С.Штильман

Поезда, пассажиры... и математика

Поезда, пассажиры... и математика

Г.А.Платонов · М.А.Файнберг · М.С.Штильман



Поезда, пассажиры... и математика

МОСКВА ТРАНСПОРТ . 1977

51:6Т1
П37
УДК 51:656.2

Scan AAW

Платонов Г. А. и др.

П37 Поезда, пассажиры... и математика. М., «Транспорт», 1977.

240 с ил. и табл. Список лит.: с. 239.

Перед загл. авт.: Г. А. Платонов, М. А. Файнберг, М. С. Штильман.

В книге в занимательной доступной форме рассказывается об основных понятиях новых областей математики, которые в последние годы нашли широкое распространение в научных исследованиях и технической литературе по железнодорожному транспорту.

Книга адресована широкому кругу инженерно-технических работников железнодорожного транспорта, она будет полезна слушателям институтов и факультетов повышения квалификации, народных университетов технического прогресса, студентам высших транспортных учебных заведений.

П $\frac{31802\text{-}189}{049(01)\text{-}77}$ 189-77 51:6Т1

© Издательство «Транспорт», 1977

ОТ АВТОРОВ

Наверное, у многих столь странное название нашей книги вызовет недоумение: как и чем они связаны между собой эти поезда, пассажиры и математика? Если говорить по существу, математизация транспортных наук — факт сегодняшнего дня. И мы, работая в этой области, постоянно сталкиваемся с необходимостью расширения сферы применения математических методов в организации деятельности железнодорожного транспорта. А где можно спокойно и обстоятельно обсудить такие проблемы? Конечно же, в уютном купе поезда дальнего следования. Здесь и попутчики могут встретиться интересные, да и время за беседой проходит быстрее.

Правда, не все у нас сразу получилось. Но полезные советы профессоров Е. С. Вентцель, И. Б. Сотникова и кандидатов наук В. Н. Воскресенского и В. А. Кудрявцева, которым мы, пользуясь случаем, приносим искреннюю благодарность, помогли в конце концов наладить разговор. Безусловно, действующие лица наших бесед — не карикатуры и не двойники живших или ныне здравствующих. Это, как писал в одной из своих книг, построенной в форме живой беседы, голландский математик А. Гейтинг, опорные точки для размещения идей и методов.

Мы просим Вас, читатель, стать участником нашего путешествия. Только непременно активным! Итак, Вы — пассажир, поезд подан, математика ждет Вас. Счастливого пути!

Приглашение к беседам

Перрон московского вокзала давно остался позади. Промелькнули дачные поселки. Жизнь в поезде постепенно входила в свое неспешное русло. Быстро познакомились. Оказалось, что все в купе ехали до конечной станции — большого сибирского промышленного и научного центра. Значит, впереди было несколько дней пути.

Математик Александр Андреевич и пожилой железнодорожник Стрелкин ехали в командировку, а студенты Галя и Борис — на свою первую производственную практику.

После того как были обсуждены проблемы необычно дождливого лета и предстоящих встреч с канадскими хоккейными профессионалами, разговор как-то сам собой прекратился и каждый из попутчиков занялся своим делом. Александр Андреевич вышел в коридор с сигаретой, Стрелкин зашуршал газетами, а студенты устроились у окна, любуясь проносящимися мимо пейзажами.

— Вот, опять пишут: «Математика помогает инженерам», — нарушил молчание Стрелкин. — В последнее время только и слышишь: математика, математика, математика. А то еще про машины электронные математические. Мой внучок Петя в четвертом классе, а туда же — математику изучает. Не то что мы в свое время арифметикой занимались. Приходит тут ко мне с задачей, а я ничего не понимаю. Неравенства, высказывания, какие-то множества. Это и есть высшая математика, что ли?

— Что вы, — засмеялась Галя, — мы с Борей высшую математику только что сдали. Но ни высказываний, ни множеств там не было. Вот только вероятности...

— Да, да, — оживился Стрелкин, — вероятности. У нас один инженер был в Москве на курсах повыше-

ния квалификации, приехал, так ни слова без этих вероятностей.

— Увлекаются сейчас математикой, очень увлекаются, — в раздумье продолжал через некоторое время Стрелкин. — Я уже почти полвека работаю на железной дороге и без всяких вероятностей, математики этой. Правда с каждым годом все меняется. Сейчас технологию все больше по науке рассчитывают. И учебники теперь не те пошли: все формулы да графики...

— Мы ведь теперь по новым программам учимся, — вступил в разговор Борис. — На последних курсах будем изучать расширенный курс математики: и теорию вероятностей, и теорию массового обслуживания, и теорию игр и математическое программирование, и...

— Опять математическое, — проворчал Стрелкин. — А родители первоклассникам задачи решить не могут! Зачем все это?

— Знаете что? — предложила Галя. — Давайте спросим об этом нашего соседа по купе Александра Андреевича. Он, как я поняла, математик, в академическом институте работает.

— Правильно! — воскликнул Борис. — Раз он математик, да еще профессионал, ему и карты в руки: пусть расскажет нам про все нововведения с математикой. Ну-ка, Галя, приглашай его в купе! Надеюсь, вы не возражаете? — обратился он к Стрелкину.

— Что вы! Ведь это я первым интерес к математике проявил.

— И нам очень интересно будет, — добавила Галя, выходя из купе. Через несколько минут она вернулась вместе с математиком.

— Александр Андреевич, — обратился к нему Стрелкин. — Мы вот тут насчет математики заговорили. Может быть, вы нам расскажете, для чего эта всеобщая

математизация понадобилась? И в школе, и в институтах, и на производстве?

— Попробую. Конечно, лучше на примере. Вот, хотя бы из вашей железнодорожной области.

— Очень интересно.

— Вам не приходилось подсчитывать частоту, ну, скажем, расцепок грузовых вагонов, поступивших на один путь, но предназначенных в разные пункты выгрузки? Нередко интересно знать, в каких комбинациях они могут находиться и сколько нужно будет сделать расцепок.

— Знаете, сам я этого никогда не делал. Но где-то читал о таком способе расчета. И поверьте, математика показалась мне здесь не очень убедительной. Ну, в простом случае все и так ясно. А попробуешь подсчитать, когда вагонов десятки и диву даешься: то частоты нужно складывать, то перемножать, а иногда и вычитать приходится. Если бы не знал, что способ наукой подтвержден, подумал, не подсчитывают, а под ответ подгоняют.

— Ну, что же, — сказал Александр Андреевич, — ваши слова мне очень кстати.

— Почему?

Попытаюсь объяснить. Дело в том, что в математике есть понятие вероятности, которое как раз и означает ожидаемую частоту наступления события. Выведены совершенно точные правила, по которым вычисляется вероятность одновременного наступления различных событий или вероятность наступления хотя бы одного из событий. Изучением этих и ряда более сложных закономерностей занимается специальная математическая дисциплина — теория вероятностей. Так вот, если бы вы в школе или в институте изучали теорию вероятностей, — хотя бы в небольшом объеме — у вас не было бы сомнений в подсчетах частоты.

Это чистая теория вероятностей и, конечно, там ничего не подгоняется под ответ, все выводится строго по правилам. Если мы будем им следовать, точно определим объем работы по расцепке вагонов. Значит, и вам, железнодорожникам нужна теория вероятностей.

— Александр Андреевич, — попросил Борис. — Не могли бы вы популярно рассказать нам, ну, скажем, о некоторых областях математики? Чем они занимаются, где и как применяются. Да так, чтобы мы и суть понять могли. А?

— Нет, так и время тратить даром не стоит, — ответил за математика Стрелкин. — Пользы от этого большой не будет. Надо не только рассказывать, но и задачки попытаться решать!

— Совершенно с вами согласен, — подтвердил Александр Андреевич. — Задачи — это, видимо, самое полезное. Но сейчас можно найти много руководств для желающих самостоятельно изучить ту или иную область математики.

— Так то книжки, — вздохнула Галя. — Книжке вопроса не задашь, а задашь, так ответа не получишь.

— А все-таки, Александр Андреевич, — предложил Борис. — Путь у нас далекий, времени много. Может быть вы расскажете нам о новых областях математики подробнее?

— Пожалуй, — после некоторого раздумья ответил Александр Андреевич, — можно попробовать. Только со временем, вы, Боря, преувеличиваете. Его в нашем распоряжении не так уж много, поэтому ограничимся разбором фундаментальных понятий. К тому же учтем и профессиональные интересы: как мне известно, на железнодорожном транспорте предметы наших бесед получили большое распространение.

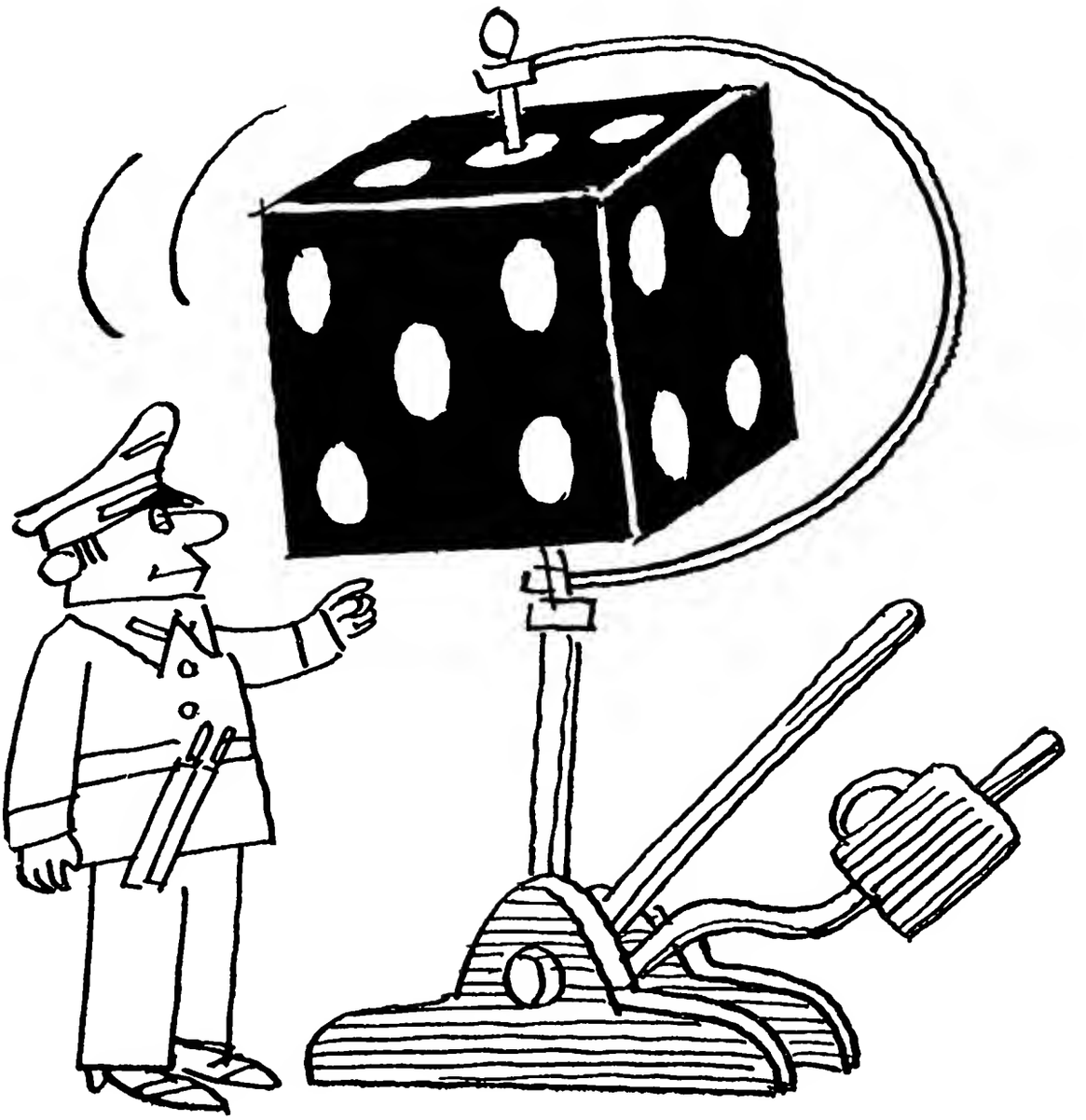
— Только должен предупредить об одном, — продолжал Александр Андреевич, — вам придется много

поработать: надо во всем попытаться разобраться и понять. Речь пойдет о новых областях математики. Для многих сами названия этих разделов связаны с чем-то непостижимым. Между тем они не слишком сложны, и овладеть их основными идеями куда проще, чем научиться решать специально подобранные замысловатые логические задачи из «Науки и жизни».

— Неужели так просто? — изумилась Галя.

— Э, нет. Так просто ничего не делается, — засмеялся Александр Андреевич, вынимая из бокового кармана блокнот.

— Вы будете еще и писать, и решать примеры, и доказывать теоремы. Но только не сегодня, конечно. Ведь недаром говорят, утро вечера мудренее...



БЕСЕДА ПЕРВАЯ

Закономерность случайностей

(Теория вероятностей)

Когда назавтра Александр Андреевич вернулся после завтрака в купе к своим любознательным слушателям, все уже ждали его.

— Прежде всего, — начал он, — я хочу задать вопрос, — обратился он к Стрелкину. — Скажите, как, по вашему мнению, мог ли я, отправляясь на завтрак в вагон-ресторан, оценить свои шансы не только вкусно поесть, но и вернуться в наш вагон к началу беседы?

— Ну, что за вопрос, конечно, могли, — нетерпеливо перебил Стрелкин. — Но при чем здесь шансы, мы хотим скорее услышать про математику.

— Не спешите, не спешите, — улыбнулся Александр Андреевич. — Это тоже про математику. Тема нашей сегодняшней беседы именно о шансах и вероятностях.

— В жизни мы часто, не отдавая себе отчета, оцениваем шансы того или иного результата опыта или наблюдения. Вот вы, Боря, когда шли за билетами на поезд, наверно, делали какие-нибудь предположения о том — достанете билет или нет.

— Еще бы! — воскликнула Галя. — Он два дня говорил, что билеты купит легко, а когда, наконец, пошел в кассу, то пришлось выстоять в очереди.

— Неправда! — перебил ее Борис. — Я не говорил наверняка, что билеты можно будет купить легко. Просто я думал, что до обычного срока отпусков еще много времени и, скорее всего, за билетами нет очередей.

— Вот видите, — сказал Александр Андреевич. — Когда Боря говорил «скорее всего», он подразумевал, что есть и другие возможности, но отдал предпочтение одной. И ошибся.

— Ну, конечно! — обиженно сказал Борис. — Я сделал только качественную оценку. Вот если бы

можно было предпочтение выразить количественно, я бы не ошибся.

— Да, было бы неплохо уметь точно оценивать шансы, — подхватил Стрелкин, — нам, железнодорожникам, тоже хотелось бы знать, когда, например, вероятнее всего начнет увеличиваться пассажиропоток. Мы могли бы заранее позаботиться о дополнительных поездах и с билетами не было бы стольких хлопот.

— Что ж, — усмехнулся Александр Андреевич, — это вполне возможно. Теория вероятностей, о которой мы уже говорили, и занимается, если можно так выразиться, количественной оценкой шансов.

— А вы могли бы рассказать нам о теории вероятностей? — попросили Галя и Борис.

— Да и я бы послушал с удовольствием, — поддержал студентов Стрелкин.

— Могу, — согласился Александр Андреевич, — но для того чтобы хорошо разобраться в основных идеях этой науки, вы должны участвовать в моем рассказе, помогать мне примерами из вашей практики.

И Александр Андреевич начал свой рассказ.

Во-первых, нужно определить предмет этой науки. Теория вероятностей изучает закономерности в случайных явлениях. При неоднократном повторении одного и того же опыта случайное явление за счет совокупного влияния многих факторов протекает каждый раз несколько по-иному. Известно, например, что при стрельбе снарядами данного типа из одного и того же орудия, установленного под определенным углом к горизонту, каждый отдельный снаряд будет иметь свою траекторию полета и будет наблюдаться рассеивание снарядов.

— Это понятно. Но как можно применять теорию вероятностей, изучающую случайные явления, к пере-

возочному процессу на железнодорожном транспорте, если перевозки планируются? — спросил Стрелкин.

— Планируется в целом объем перевозок за определенный период. Но в процессе выполнения плана объем перевозок в силу многих влияющих факторов колеблется и не является одинаковым в различные сутки, декады и месяцы. Поэтому и возможно при расчете, например, мощности технических средств железнодорожного транспорта применять те или иные положения теории вероятностей. Да вы и сами говорили, что железнодорожникам важно знать, когда начнет увеличиваться пассажиропоток, — обратился Борис к Стрелкину.

— Для изучения случайных явлений, — продолжил Александр Андреевич, — в теории вероятностей имеется ряд основных понятий. Важнейшим из них является событие, подразумевающее всякий факт, который в результате опыта может произойти или не произойти.

Например, событие, состоящее в том, что завтра будет солнечный день, или событие, состоящее в том, что при бросании игрального кубика во многих детских настольных играх выпадает четное число очков. Для того чтобы оценивать шансы на появление этих или каких-нибудь других событий, надо сначала научиться оперировать с самими событиями. Действия над событиями определяются специальными законами.

Среди всех событий, которые могут произойти в результате опыта или наблюдения, выделяют два особых. Это так называемые невозможное и достоверное события. Невозможное — это такое событие, которое не может появиться в результате нашего опыта. Например, событие, заключающееся в том, что завтра будет безоблачный день и одновременно будет идти про-

ливной дождь, явно невозможно. Так же невозможно событие, заключающееся в том, что в результате одного бросания игрального кубика выпадет 13 очков. Достоверное событие — такое, которое произойдет непременно. Таким является событие, заключающееся, например, в том, что в результате бросания монеты она упадет на стол так, что мы увидим либо герб, либо ее номинал, или, как говорят, — герб или решетку.

— А вдруг монета станет на ребро? — спросила Галя.

— Такого не может быть, — снисходительно возразил Борис.

— Нет. Это не такой простой вопрос, — сказал Александр Андреевич. — Здесь я действительно упустил один момент. Мы к нему вернемся позже. Какое же придумать достоверное событие?

— А вот такое событие, — вмешался Стрелкин. — Допустим, что на станцию прибыл поезд. Тогда мы можем утверждать, что при расформировании его состава вагоны будут направлены не менее чем на два пути в сортировочном парке.

— Что же, — ответил Александр Андреевич, — это действительно достоверное событие. Для того чтобы нам было удобнее разговаривать в дальнейшем, договоримся обозначать невозможное событие буквой V, а достоверное — буквой U.

— Но ведь есть много невозможных и достоверных событий. Что же, все они будут обозначены одной и той же буквой? — удивился Борис.

— Как мы увидим, это не приведет к недоразумению, потому что мы будем рассматривать только события, могущие произойти в результате какого-нибудь вполне определенного эксперимента, и в этом случае невозможное и достоверное события будут иметь однозначный смысл.

Рассмотрим два события — A и B. Произведением AB этих событий назовем событие, состоящее в том, что произошли одновременно и A и B, а суммой $A+B$ — событие, состоящее в том, что произошло хотя бы одно из этих событий. Если, например, A — событие, заключающееся в том, что при бросании игрального кубика выпадает четное число очков, а B — в том, что выпадает число очков, делящееся на три, то событие AB есть, очевидно, просто выпадание 6 очков.

— А сумма тогда есть выпадание любого числа очков, кроме 1 и 5. Да? — спросила Галя.

— Совершенно верно, — ответил Александр Андреевич.

— Я не совсем понял. Можно объяснить этот пример подробней? — попросил Стрелкин.

— Ну, как же, — вмешался Борис, — если A — это выпадание четного числа очков, то A произойдет, если выпадает одно из трех чисел — 2, 4 или 6. Точно так же B произойдет, если выпадает одно из двух чисел — 3 или 6.

— Прекрасно, — сказал Александр Андреевич. — Может быть, вы мне ответите, что это за события — AV, AU, $A+V$, $A+U$?

— Попробую. А что это за событие — A?

— Любое. Какое-то событие, появляющееся в результате эксперимента.

— Значит так. AV — это событие, состоящее в том, что произошли одновременно событие A и невозможное событие. Да, но тогда оно само невозможно! Значит $AV=V$?

— Верно, — подбодрил Стрелкина математик.

— Так, пойдем дальше, — уже увереннее продолжил Стрелкин. — AU — событие, состоящее в том, что произошли одновременно события A и достоверное событие. Но достоверное событие всегда происходит.

Значит надо, чтобы просто произошло событие A, т. е. $AU=A$. Событие $A+V$ произойдет тогда, когда произойдет либо A, либо невозможное событие. Невозможное событие произойти не может, значит $A+V=A$. Понятно, что $A+U=U$, потому что достоверное событие выполняется всегда, а значит и $A+U$ выполняется.

— Очень хорошо, — сказал Александр Андреевич, — вы, наверно, согласны, что $AA=A$ и $A+A=A$.

— Ну, конечно же, это ясно, — в один голос ответили слушатели.

— Пойдем дальше, — продолжил Александр Андреевич. — Аналогично сумме и произведению двух событий можно определить сумму $A+B+C=(A+B)+C$ и произведение $ABC=(AB)\,C$ трех и вообще любого числа событий. Понятно, что эти события не зависят от того, в каком порядке мы будем брать их составные части, т. е.

$$(A+B)+C=A+(B+C)=A+B+C=A+C+B=\ldots$$

$$\text{и}\quad (AB)C=A(BC)=ABC=ACB=\ldots$$

Если произведение двух событий A и B есть невозможное событие, т. е. если они не могут произойти одновременно, то они называются несовместными.

Если имеется несколько событий A_1, A_2, A_3,... A_n, которые не могут произойти одновременно, т. е.

$$A_1A_2\ \ldots\ A_n=V,$$

то они называются несовместными в совокупности. Если среди группы событий есть хотя бы одна пара несовместных, то вся эта группа несовместна в совокупности, так как произведение этой пары уже дает невозможное событие. Но может оказаться, что среди выбранной группы событий нет ни одной несовместной пары, а в совокупности они несовместны.

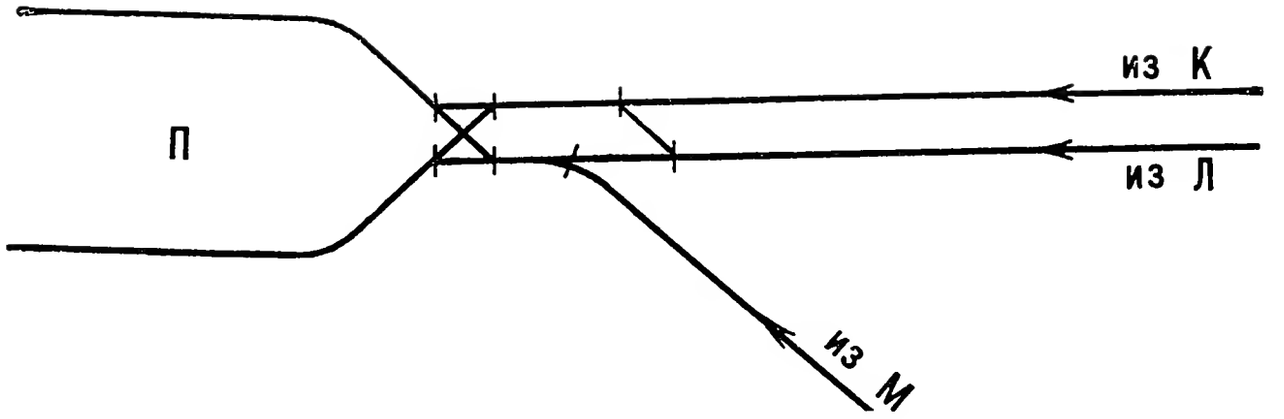


Рис. 1.1.

— Как же это может быть? — удивился Стрелкин.

— Очень просто, — сказал Борис, — и, быстро начертив схему станции (рис. 1.1), предложил обозначить через A событие, заключающееся в приеме в парк $П$ поезда, прибывающего из $К$, через B — поезда, прибывающего в этот парк из $Л$, и через C — поезда, прибывающего из $М$.

— Тогда, — продолжил Борис, — возможны произведения событий AB, AC и BC, т. е. среди событий A, B и C нет ни одной несовместной пары (возможен одновременный прием из $К$ и $Л$, из $К$ и $М$ и из $Л$ и $М$). В совокупности же все три события (произведение событий ABC) несовместны, так как одновременный прием поездов со всех трех направлений невозможен.

— Продолжим, — сказал Александр Андреевич. — Пусть имеются попарно несовместные события A_1, A_2, ..., A_n. Тогда событие $A_1+A_2+...+A_n$ состоит в том, что произошло хотя бы одно из них (и поскольку никакие два из них не могут произойти одновременно, то это событие означает, что произошло только одно из них).

Если $A_1+A_2+...+A_n=U$, т. е какое-то из этих событий произошло обязательно, то эти события образуют полную группу событий.

Мы определили умножение и сложение событий. Теперь выясним, как связаны эти две операции.

Возьмем три события A, B и C и рассмотрим событие $A\ (B+C)$. Что это за событие? Ну-ка, Галя.

— Это значит, что произошло событие A и хотя бы одно из событий B или C.

— Правильно, — подхватил Борис, — или AB или AC, а это есть событие $AB+AC$.

— Прямо как в алгебре, — удивился Стрелкин, —

$$A(B+C)=AB+AC.$$

— Совершенно верно. Потому и множество событий вместе с введенными операциями называют алгеброй событий. Но если есть сложение, очевидно, должно быть и вычитание. К сожалению, здесь все происходит не так, как в обычной алгебре. Впрочем, уже равенство $A+A=A$ было необычным.

Определим сначала противоположное событие к событию A. Его обозначим через $\overline{A}$. Это событие происходит тогда, когда не происходит A, и наоборот.

Из определения понятно, что если мы возьмем противоположное событие к $\overline{A}$, то получим само событие A. Значит, события A и $\overline{A}$ можно называть противоположными. Такими, например, являются события U и V. Поэтому $\overline{U}=V$ и $\overline{V}=U$.

Вы легко проверите, что $A\overline{A}=V$ и $A+\overline{A}=U$, т. е. любая пара событий A и $\overline{A}$ образует полную группу.

Далее рассмотрим два события A и B. Разностью $A-B$ или дополнением события B в событии A называется событие, состоящее в том, что происходит событие A и не происходит событие B.

Разность $A-B$ не является операцией обратной сложению, потому что $(A-B)+B$ не равно A.

Разность $A-B$ можно еще записать в виде $A\overline{B}$. Потому что и то и другое событие означает, что одновременно произошло событие A и не произошло событие B.

Говорят также, что событие A содержится в событии B, если наступление события A всегда влечет за собой появление события B. Обозначать мы это будем так: $A \subseteq B$.

Вы, наверно, легко разберетесь, что для любых двух событий A и B $AB \subseteq A$, так как слева стоит событие, для выполнения которого необходимо выполнение события A. Кроме того, для любых событий A и B таких, что $A \subseteq B$, справедливо соотношение $\overline{B} \subseteq \overline{A}$. Действительно, пусть произошло событие $\overline{B}$, т. е. не произошло событие B. Если бы событие A произошло, то из $A \subseteq B$ следовало бы, что и B произошло. Значит из того что произошло $\overline{B}$, следует, что A не произошло, т. е. произошло $\overline{A}$.

Так как достоверное событие имеет место всегда, то для любого события A $A \subseteq U$. Для единства всех соотношений считают, что для любого A $V \subseteq A$. Если события A и B связаны соотношением $B \subseteq A$, то $AB = B$ и $A + B = A$.

— Да-а, — протянул Стрелкин, — эти формулы меня, признаться, уже утомили.

— Что же, — сказал Александр Андреевич — тогда мы можем от них отойти и вернуться к опыту с бросанием монетки или кубика. Вопрос, который нам предстоит разрешить, состоит в следующем. Какие события нам нужно рассматривать?

Этот вопрос относится не столько к математике, сколько к содержанию эксперимента, который мы производим. Мы вольны, конечно, рассматривать все мыслимые события, которые могут произойти в результате опыта или наблюдения.

Например, в эксперименте с бросанием игрального кубика в качестве результата можно принимать во внимание не только количество выпавших очков, но и чис-

ло оборотов, которые кубик сделает в воздухе. Или интересоваться только тем, четно выпавшее количество очков или нечетно.

Но при попытке учитывать все явления, сопутствующие эксперименту, станет чрезвычайно трудным описать все возможные исходы. Если же все исходы делить на слишком крупные группы, будет невозможно определить, произошло или нет какое-нибудь событие.

Например, если исходами бросания кубика считать только четность числа выпавших очков, то нельзя сказать, произошло ли событие, состоящее в том, что число очков делится на 3. Поэтому задача состоит в том, чтобы каждому эксперименту сопоставить модель, в которой будут описаны все возможные исходы. Их мы будем называть элементарными. Таким образом, каждое интересующее нас событие можно разложить на эти элементарные исходы.

В случае бросания монеты мы можем пренебречь ее толщиной и возможностью упасть на ребро и считать, что есть всего два элементарных исхода — выпадание герба и выпадание решетки. В случае с кубиком элементарными исходами можно считать выпадания различных чисел.

Системой или, как принято говорить, пространством элементарных исходов является любая полная группа попарно несовместных событий. Нужно только, чтобы все интересующие нас события можно было разложить на элементарные.

Рассмотрим такую ситуацию. Пусть в депо на профилактический осмотр прибыли три электровоза. Осмотр производят две бригады: одна проверяет электрическую часть, а другая — механическую. Предположим, что направления на электровоз бригады получают в разных местах и независимо друг от друга. Как выбрать пространство элементарных исходов?

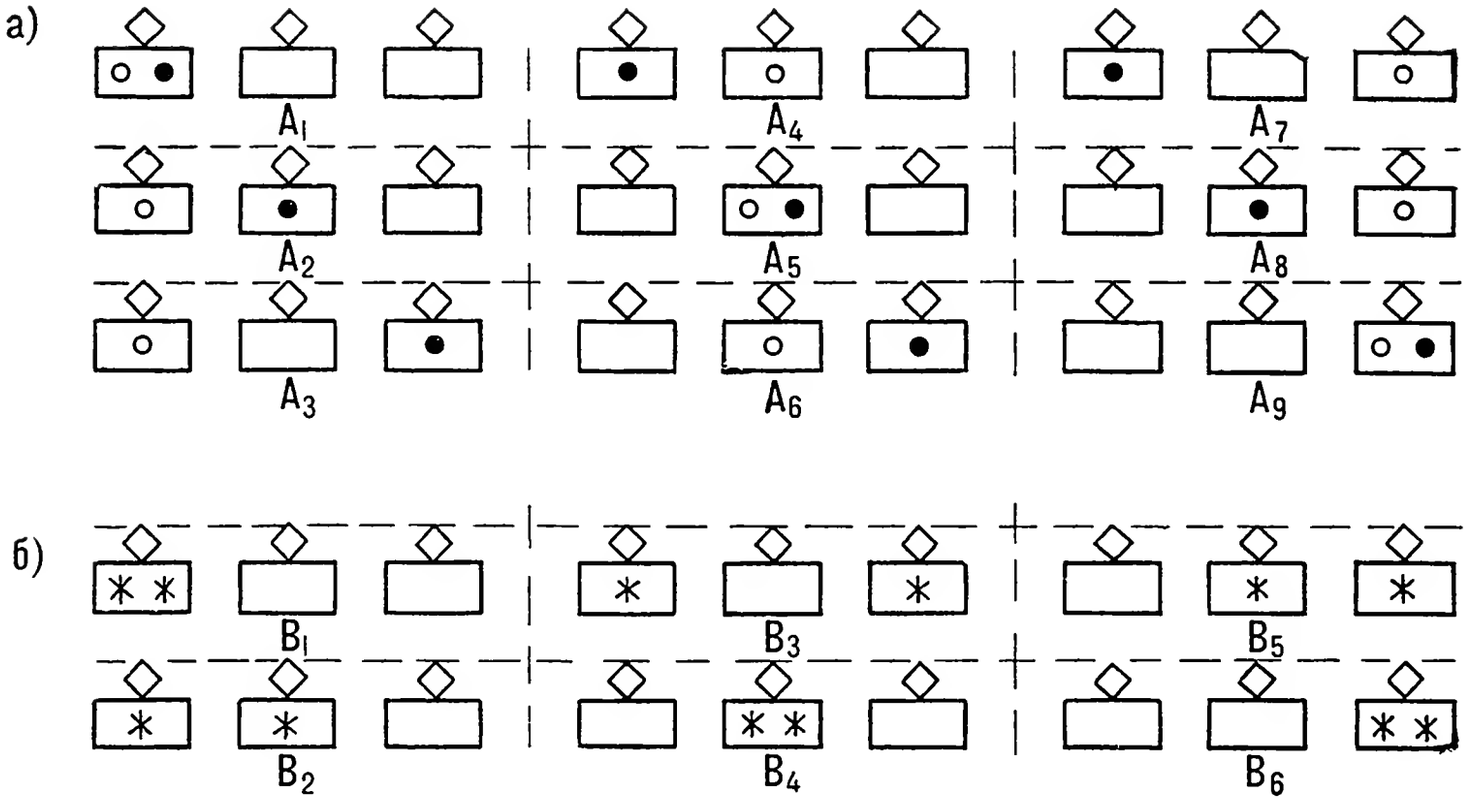


Рис. 1.2

— Самое естественное, — сказал Стрелкин, — будет выявить все возможные распределения бригад по электровозам. Если, например, бригаду электриков обозначить белым кружком, а бригаду механиков черным, то все варианты можно выписать так (см. рис. 1.2, *а*).

Здесь событие A_2 означает, что электрики работают на 1-м, а механики на 2-м электровозах, а событие A_9 — то, что обе бригады на 3-м электровозе.

— Правильно, — подтвердил Александр Андреевич.— Но как определить, все ли варианты мы перебрали?

— Ну, это совсем просто, — сказал Борис. — Бригада электриков может быть направлена на любой из трех электровозов. При каждом направлении этой бригады у бригады механиков есть тоже три возможности. Значит, всего различных вариантов ровно $3 \times 3 = 9$. А так как среди выписанных нами случаев нет одинаковых, то они и описывают все возможные распределения бригад по электровозам.

— Ну что ж, верное рассуждение. Обратимся к этим выписанным событиям. Они очевидным образом попарно несовместны, так как никакие два из них не могут произойти одновременно, и в сумме составляют достоверное событие. Значит, они образуют действительно полную группу и могут служить пространством элементарных исходов. Попробуем выразить через эти события какие-нибудь другие. Например, как записать событие A, состоящее в том, что бригада электриков работает на 2-м электровозе?

— Наверно, — начала Галя, — это будет событие A_4.

— Или A_5, — подхватил Стрелкин.

— Но A_6 тоже означает, что электрики на 2-м электровозе, — возразил Борис.

— Вы все, конечно, правы, — сказал Александр Андреевич. — Появление каждого из этих событий влечет за собой событие A, т. е., как мы говорим, каждое из событий A_4, A_5 и A_6 содержится в A. С другой стороны, если произошло событие A, то возможны три случая: бригада механиков работает на 1-м, на 2-м или на 3-м электровозе. В 1-м случае выполняется событие A_4, во 2-м — A_5 и в 3-м — A_6. Значит, событие выполняется тогда и только тогда, когда выполнено какое-нибудь из этих трех событий, т. е. когда произошло событие $A_4+A_5+A_6$. Это записывается так:

$$A = A_4 + A_5 + A_6 .$$

— А теперь, как записать событие, состоящее в том, что обе бригады оказались на одном электровозе?

— Ну, это просто, — ответила Галя, — надо сложить события A_1, A_5 и A_9.

— Хорошо, здесь, кажется, мы разобрались, — продолжил Александр Андреевич. — Попробуем немножко изменить ситуацию. Предположим, что в депо попал неспециалист, который не может сразу определить, какая

бригада перед ним — электрики или механики. Он только различает, что бригады две. Как для него будет выглядеть пространство элементарных исходов?

— Наверно, он некоторые из прежних различных исходов будет путать, — сообразил Боря.

— Правильно, — подтвердил Александр Андреевич. — Какие же?

— Ну, например, A_2 и A_4 или A_6 и A_8.

— Верно. Выпишем это пространство. Теперь бригады будем обозначать звездочкой* (см. рис. 1.2, *б*).

Здесь, как легко видеть, события B_1, B_2, B_3, B_4, B_5 и B_6 также образуют полную систему. Но уже событие A, состоящее в том, что электрики работают на 2-м электровозе, мы через них не выразим.

— Что ж это за пространство элементарных исходов, если мы не можем выразить через них такое простое событие? — усмехнулась Галя.

— Это объяснить нетрудно, — возразил Стрелкин. — Вот если я, например, планирую работу бригад, то мне все равно, какая где работает. Мне важно, чтобы бригады не попали на один электровоз и не мешали друг другу.

— Прекрасно! — обрадованно воскликнул Александр Андреевич. — А ведь событие, состоящее в том, что обе бригады попали на один электровоз, выражается и в новом пространстве элементарных событий, а именно $B_1+B_4+B_6$.

Итак, мы научились выделять все возможные элементарные исходы какого-нибудь эксперимента и выражать через них все интересующие нас события. Теперь мы переходим к самому интересному: как оценивать шансы на появление того или иного события?

Пусть пространство элементарных исходов состоит из конечного числа событий A_1, A_2, ..., A_n, как это было у нас до сих пор. Мы будем говорить, что на простран-

стве элементарных исходов $A_1, A_2, ..., A_n$ заданы вероятности, если каждому исходу A_i сопоставлено неотрицательное число $p_i \geqslant 0$ и все эти числа в сумме дают единицу: $p_1+p_2+...+p_n=1$. Эти числа мы и будем считать вероятностями элементарных исходов, записывая это так: $P\{A_i\}=p_i$.

Если теперь нам понадобится определить вероятность какого-нибудь составного события, то мы возьмем и сложим вероятности событий, составляющих его. Например,

$$P\{A_1+A_3+A_{n-1}\}=p_1+p_3+p_{n-1}.$$

— Ну, вот, приехали! — разочарованно протянула Галя. — А как же определить вероятности элементарных исходов? Опять интуиция?

— Да, это самое тонкое место в обосновании теории вероятностей, — ответил Александр Андреевич. — Конечно, определение вероятностей элементарных исходов связано с интуитивными соображениями, но для того, чтобы субъективность интуиции не свела все усилия на нет, надо было научиться объективно оценивать правильность интуитивных представлений. И теория вероятностей справилась с такой задачей. Мы еще будем говорить об этом.

А сейчас давайте рассмотрим опять наш игральный кубик. Предположим, что кубик новенький, сделанный на самых точных станках. Можем ли мы отдать предпочтение одному какому-нибудь из шести возможных исходов бросания?

— Нет, наверно, — неуверенно произнес Борис. — Раз он сделан абсолютно симметрично, то выпадания любых граней равновозможны.

— Что ж, это и есть некое интуитивное соображение, подсказывающее нам, что всем возможным исходам бросания надо приписать одинаковую вероятность.

А так как всего исходов шесть и суммарная вероятность равна 1, каждому исходу следует приписать вероятность по $\frac{1}{6}$.

Обратимся теперь к более сложному примеру с ремонтными бригадами. Здесь девять событий. Можно ли считать их равновозможными?

— А почему бы и нет, — сказала Галя. — Ведь бригады назначаются независимо друг от друга, и у каждой из них одинаковые шансы попасть на любой электровоз. Значит, мы можем считать все исходы равновозможными и приписать каждому из них вероятность $\frac{1}{9}$.

— А как же тогда быть в случае неразличимых бригад? — спросил Борис. — Ведь если так, то, с одной стороны, мы должны всем событиям и, в частности, B_2 приписать вероятность $\frac{1}{6}$, а с другой стороны, $B_2 = A_2 + A_4$ и

$$P\{B_2\} = P\{A_2\} + P\{A_4\} = \frac{1}{9} + \frac{1}{9} = \frac{2}{9} \neq \frac{1}{6}.$$

— Как это так? Одно и то же событие имеет, оказывается, две различные вероятности.

— Вы, Боря, допустили ошибку. Дело в том, что вы вычисляли вероятность события B_2 в разных пространствах элементарных исходов. Если вы считаете, что невозможность различить бригады не сказывается на существе дела, то вы не имеете права считать все исходы B_1, B_2, ..., B_6 равновероятными, а должны приписать им вероятности в соответствии с правилом вычисления вероятности составного события, считая события $B_2 = A_2 + A_4$, $B_3 = A_3 + A_7$ и $B_5 = A_6 + A_8$ составными в пространстве исходов A_1, A_2, ..., A_9. Если же вы считаете события B_1, B_2 ..., B_6 равновозможными, тем са-

мым вы подразумеваете, что неразличимость бригад существенна и носит объективный характер. Это и есть проблема о критериях правильности определения вероятностей элементарных исходов. Мы займемся ею позже.

А сейчас рассмотрим пространство равновозможных элементарных исходов. Допустим, что имеется n элементарных исходов и, следовательно, каждый из них имеет вероятность $\frac{1}{n}$. Пусть событие A представляет собой сумму k различных элементарных исходов. Тогда вероятность этого события A будет, по определению, равна сумме k одинаковых слагаемых по $\frac{1}{n}$, т. е. равна $\frac{k}{n}$. Если элементарные исходы, составляющие A, назвать благоприятными, вероятность события A оказывается равной отношению числа благоприятных исходов к общему числу всех равновозможных исходов. Мы пришли к так называемому классическому определению вероятности события.

— Как просто! А нельзя всегда выбирать элементарные исходы равновозможными? — спросил Стрелкин.

— К сожалению, это не всегда удается. Но, конечно, к этому нужно стремиться. Даже если приходится для этого брать в качестве элементарных исходов «слишком» мелкие. К тому же интуитивно легче определить равновозможность двух событий, чем во сколько раз одно вероятнее другого. Каковы свойства вероятностей? Во-первых, поскольку достоверное событие — сумма всех элементарных, его вероятность равна 1.

Во-вторых, подумайте, пожалуйста, и скажите, как определить вероятность того, что наступит хотя бы одно из двух произвольных событий A и B, т. е. чему равна вероятность события $A+B$?

— Можно мне? — сразу заторопился Борис. — Надо действовать по определению. Взять элементарные исхо-

ды, составляющие событие A, элементарные исходы, составляющие событие B, и сложить их вероятности.

— Да, но может оказаться, что один и тот же элементарный исход входит и в A и в B. Ведь эти события не обязательно несовместны, — возразила Галя.

— Правильно, — ответил ей Борис, — если какой-то элементарный исход входит и в A и в B, то его вероятность надо прибавлять к общей сумме не два раза, а один.

— Подождите, но насколько я понял, это значит, что вероятность $A+B$ не равна сумме вероятностей A и B? — вмешался Стрелкин.

— Совершенно верно, — ответил Александр Андреевич. — Если элементарный исход входит и в A и в B, то в сумму, составляющую $P\{A+B\}$, его вероятность входит один раз, а в $P\{A\}+P\{B\}$ — два раза. Значит, $P\{A+B\} \leqslant P\{A\}+P\{B\}$, причем правая часть больше левой на величину, равную сумме вероятностей исходов, входящих как в A, так и в B, т. е. исходов, составляющих событие AB. Поэтому можно записать

$$P\{A+B\}=P\{A\}+P\{B\}-P\{AB\}. \qquad (1.1)$$

Если события A и B несовместны, т. е. $AB=V$ (A и B не содержат общих исходов), то

$$P\{A+B\}=P\{A\}+P\{B\}. \qquad (1.2)$$

Так, например, $P\{A\}+P\{\bar{A}\}=P\{A+\bar{A}\}=P\{U\}=1$. В частности, если $A=U$ и $\bar{A}=\bar{U}=V$, то $P\{U\}+P\{V\}=1$ и $P\{V\}=1-P\{U\}=1-1=0$.

Формула (1.1) для произвольной пары событий и ее частный случай (1.2) для несовместных событий выражают теорему сложения вероятностей.

— Вот это да! — воскликнула Галя. — Мы, оказывается, доказали теорему!

— Верно, — усмехнулся Александр Андреевич, — и нам еще предстоит доказать несколько других. А над чем это вы, Боря, задумались?

— Меня все-таки интересует такая вещь, — ответил Борис. — Если мы бросаем кубик, то вероятность выпадания числа, делящегося на 1, равна $\frac{1}{6}$. Но предположим, что кубик бросала Галя, а мне сказала только, что выпало четное число очков. Тогда я могу рассуждать так: раз выпало четное число очков, значит, это либо 2, либо 4, либо 6 — и все исходы равновозможны. Из них на 4 делится только 4. Значит, вероятность того, что выпавшее число делится на 4, равна $\frac{1}{3}$. Как же так?

— Все правильно, — сказал Александр Андреевич.— Только $\frac{1}{6}$ — это вероятность выпадания числа, делящегося на 3 без дополнительных условий, а $\frac{1}{3}$ — вероятность того же события при условии, что выпавшее число четно.

Здесь мы столкнулись с так называемой условной вероятностью. Давайте рассмотрим еще пример. Предположим, что на двух станках обрабатываются какие-нибудь детали. На одном работает опытный мастер, который допускает брак только в 5 случаях из 100, а на другом — молодой ученик, который портит каждую пятую деталь.

Возьмем 100 деталей, сделанных на одном станке, и 100 деталей, сделанных на другом. А потом из полученной партии в 200 деталей выберем наугад одну. Какова вероятность события B, состоящего в том, что выбранная деталь сделана мастером?

— Наверно $\frac{1}{2}$, — сказал Стрелкин.

— Правильно. Всего среди 200 деталей 100, сделанных мастером, а все исходы — выбор каждой детали — равновозможны.

Пусть теперь выбранная нами деталь оказалась бракованной (это событие обозначим через A). Какова вероятность того, что она была сделана мастером (событие B)?

— Это задача потруднее, — задумался Стрелкин.

— Ну почему же? — возразил Александр Андреевич. — Давайте рассуждать, как и раньше. Всего элементарных исходов 200, и все они равновозможны. Сколько имеется элементарных исходов, составляющих событие A?

— Это понятно, 5 деталей мастера и 20 — ученика, — быстро ответил Стрелкин, — всего, значит, 25.

— Значит, — продолжил Александр Андреевич, — известие о том, что выбранная деталь бракованная, ограничивает исходы до 25, и все они равновозможны. Мы можем принять их за новое пространство элементарных исходов. Сколько же из них нас устраивает?

— Понятно, что устраивают нас те из исходов, составляющих A, которые входят в событие B, т. е. бракованные детали, сделанные мастером, — 5 штук.

— Значит, для того чтобы получить нужную вероятность, — сказала Галя, — следует 5 поделить на 25. Правильно?

— Правильно, — ответил Александр Андреевич. — Эта условная вероятность события B при условии события A, которую мы обозначим $P\{B/A\}$, равна $\frac{5}{25}=\frac{1}{5}$. Но давайте теперь внимательно проследим, что же мы сделали. Из общего числа исходов n (в нашем случае их было 200) мы выделили исходы, благоприятные для со-

бытия A. Их оказалось $n_A=25$. При указанном условии эти исходы составили новую полную систему возможных исходов, так как событие A стало достоверным. Затем мы выбрали исходы, входящие в A и благоприятные для B, т. е. исходы, благоприятные для события AB. Их оказалось $n_{AB}=5$. После этого мы вычислили вероятность по правилу отношения числа благоприятных исходов к общему числу исходов: $P\{B/A\}=\frac{n_{AB}}{n_A}$.

Для того чтобы вычислять условные вероятности не только в случае пространств равновозможных элементарных исходов, мы нашу формулу немного изменим. Замечая, что $\frac{n_{AB}}{n}=P\{AB\}$, а $\frac{n_A}{n}=P\{A\}$ (здесь участвуют безусловные вероятности), получим следующее соотношение для условных вероятностей

$$P\{B/A\}=\frac{n_{AB}}{n_A}=\frac{P\{AB\}}{P\{A\}}. \qquad (1.3)$$

Теперь уже не имеет значения, как заданы вероятности элементарных исходов — равными или нет. Как видно из этой формулы, условные вероятности остались неопределенными, если $P\{A\}=0$.

— А почему, когда вы говорили о введении вероятностей элементарных исходов, допускали, что они могут быть равны 0? — спросил Борис.

— А почему бы и нет? — удивился математик.

— Зачем тогда вообще рассматривать такие элементарные исходы? Ведь если какому-нибудь исходу приписали вероятность 0, то это означает, что он практически не может произойти, т. е. этот исход практически невозможен.

— Видите, Боря, — вы сами говорите, что этот исход не «невозможен», а «практически невозможен», т. е. мысленно вы такой исход можете предположить, ну, на-

пример, возможность монеты упасть на ребро. А вот «практическую невозможность» какого-либо события надо еще подтвердить экспериментально. Все опять сводится к проверке правильности установленных нами вероятностей элементарных событий.

— Но если некоторые элементарные события имеют нулевую вероятность, значит из равенства $P\{A\}=0$ вовсе не следует, что событие A невозможно. Верно? — заметил Стрелкин.

— Конечно! — подтвердил Александр Андреевич. — И это нужно хорошо понимать. Дело в том, что мы до сих пор рассматривали только пространства, состоящие из конечного числа элементарных исходов.

А вот рассмотрим, например, такую ситуацию. Едем мы сейчас в поезде, который ведет электровоз. И вдруг поезд останавливается. Все волнуются — в чем дело? Оказывается, что где-то на участке, по которому идет наш поезд, произошел обрыв контактного провода. Причем неизвестно где. Какова вероятность того, что разрыв произошел перед поездом? Как вы думаете?

— Здесь в качестве равновозможных элементарных исходов можно, наверно, взять различные возможные места разрыва провода, — начал Борис.

— Да, конечно, — сказал Александр Андреевич, — но вот беда, таких мест бесконечно много — целый отрезок.

— Их даже пересчитать нельзя, — вставила Галя.

— Да, верно, — подтвердил Александр Андреевич.— Это случай, который мы будем называть непрерывным пространством элементарных исходов в отличие эт случаев, когда элементарных исходов конечное число или бесконечное, но все же такое, что их можно пересчитать, т. е. перенумеровать.

— А разве нельзя взять и перенумеровать все точки отрезка? — удивился Стрелкин. — Возьмем одну точку, назовем ее первой, потом вторую и т. д.

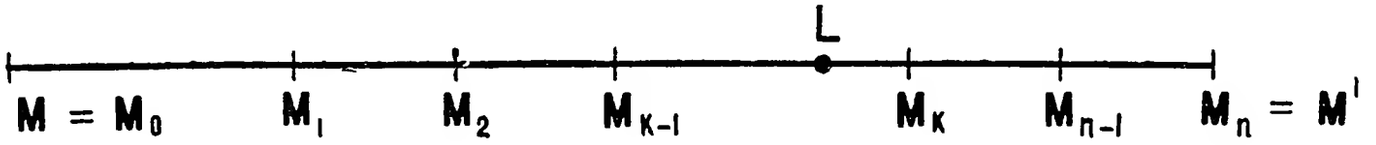
— Ну нет! — возразил Борис, — нам на лекциях доказывали, что так не перебрать всех точек отрезка.

— Совершенно верно, — продолжил Александр Андреевич. — Точки отрезка перенумеровать нельзя. Но вернемся к вопросу, с которого начали. Какие же вероятности приписать нашим элементарным исходам?

— Если их бесконечно много и все они равновозможны, то остается только приписать им нулевые вероятности, — неуверенно сказал Борис. — Ведь если у них будут положительные вероятности, то сумма будет бесконечной, а должна быть единица.

— Да, но если вероятности всех элементарных исходов — нули, то и любое составное событие будет иметь нулевую вероятность. И вероятность обрыва провода вообще, а не только перед поездом будет равна нулю, — сказала Галя.

— Да, значит, что-то здесь не совсем правильно, — задумчиво произнес Борис. — А если попробовать так, возьмем весь участок, на котором произошел обрыв, и разобьем его на равные отрезки. Пусть это будут точки $M_0=M$ — начало участка, M_1, M_2, ..., $M_n=M'$ — конец участка:



Тогда, раз все отрезки равной длины, события A_k, состоящие в том, что разрыв произошел на отрезке $M_{k-1}\,M_k$ $(k=1, 2,...)$, равновозможны. Все они попарно несовместны и в сумме составляют достоверное событие, заключающееся в том, что разрыв произошел на всем участке MM_n. Значит, они образуют пространство элементарных исходов с равными вероятностями — по $\frac{1}{n}$.

— А чтобы подсчитать вероятность того, что разрыв произошел перед поездом, остановившимся в точке L, надо подсчитать количество маленьких отрезков, уместившихся на участке LM', и поделить их на n, — закончил Борис радостно.

— Здорово! — воскликнула Галя и захлопала в ладоши. — Молодец, Борька!

— Здорово, да не совсем, — усмехнулся Александр Андреевич. — А если точка L попадает не на границу между двумя отрезками, а внутрь какого-нибудь отрезка, например, $M_{k-1}M_k$. Как тогда считать вероятность $\frac{n-k}{n}$ или $\frac{n-k+1}{n}$?

— Ну, если n очень большое, — ответил Борис, — то $\frac{n-k+1}{n} - \frac{n-k}{n} = \frac{1}{n}$ — очень маленькая величина, и эти числа почти одинаковы. Поэтому любое из них можно считать приблизительно равным настоящей вероятности.

— Что же, для практического вычисления эти соображения можно, конечно, принять. Но ведь нас сейчас интересует теория, так сказать, идейная сторона вопроса, и поэтому мы не должны пренебрегать такими «мелочами», как $1/n$. А уж если мы ими пренебрегаем, то надо это обосновывать. Почему эти числа близки к «настоящей» вероятности? Откуда такая уверенность? Ведь мы еще не знаем, чему она равна — эта «настоящая» вероятность. Однако в рассуждениях Бориса есть идея, которая приведет нас к нужному ответу.

Возьмем на отрезке MM' два произвольных, но равных по длине отрезка N_1N_1' и N_2N_2'

События, состоящие в том, что разрыв произошел внутри одного из них или другого, по-видимому, разумно считать равновозможными — по тем же причинам, по которым Боря считал равновозможными разрывы на каждом из маленьких отрезков — потому, что они имеют одинаковую длину. Иными словами, естественно считать, что чем больше отрезок, тем больше вероятность того, что разрыв произошел именно на нем, и совершенно неважно, в начале, в середине или в конце всего участка MM' этот отрезок находится. Если мы увеличим отрезок вдвое, то естественно считать, что и вероятность разрыва на нем увеличилась вдвое. Таким образом, мы приходим к такому предположению: вероятность того, что разрыв произошел на данном отрезке, пропорциональна его длине.

Поэтому, если событие A состоит в том, что разрыв произошел на отрезке LM', вычислить его вероятность можно так:

$$P\{A\} = \frac{\text{длина } LM'}{\text{длина } MM'}.$$

Отсюда, кстати, видно, что если в схеме Бори увеличивать n, то длина M_kM' будет приближаться к длине LM', а длина $M_k M'$ равна $(n-k)\times$(длина маленького отрезка) $= (n-k) \times \frac{\text{длина } MM'}{n} = \frac{n-k}{n} \times$ (длина MM') и $\frac{n-k}{n} = \frac{\text{длина } M_k M'}{\text{длина } MM'}$, т. е. с ростом n действительно приближается к $P\{A\}$. Ну, и вероятность «элементарного исхода» — попадания в какую-нибудь точку — действительно равна нулю, так как «длина» точки — нуль.

Здесь нам пришлось определить вероятность способом, отличным от того, которым мы пользовались, когда вводили вероятность на конечном пространстве элементарных исходов. Мы еще вернемся к этому вопросу

и увидим, что эти способы на самом деле имеют общую основу.

Надеюсь, вы не забыли формулу (1.3), ее можно понимать как определение условной вероятности, подразумевая, что вероятность условия $P\{A\}$ больше нуля.

Если мы будем рассматривать условные вероятности различных событий при одном и том же условии C, то можем выбрать новое пространство элементарных исходов — исходы, составляющие событие C, а вместо события B рассматривать событие CB. Нужно только все вероятности умножить на один и тот же множитель $\frac{1}{P\{C\}}$, чтобы сделать суммарную вероятность всех элементарных исходов равной 1. Поэтому для условных вероятностей справедливы такие же соотношения, как и для безусловных. В частности, теорема сложения [см. формулу (1.1)] приобретает вид:

$$P\{A+B/C\}=P\{A/C\}+P\{B/C\}-P\{AB/C\}. \quad (1.4)$$

Формулу (1.3) можно переписать так:

$$P\{AB\}=P\{A\}\,P\{B/A\}. \quad (1.5)$$

Это есть теорема умножения вероятностей. События A и B можно поменять местами, если $P\{B\}$ тоже отлична от нуля. Получим

$$P\{AB\}=P\{B\}\,P\{A/B\}.$$

— Еще одна теорема? — спросила Галя. — Такая простая?

— Что значит простая? — переспросил Александр Андреевич.

— Ну как же, — объяснила Галя. — Взяли определение, переписали его по-другому и получили теорему. Хорошо бы все теоремы так доказывались.

— Нет, здесь не все так просто, — улыбнулся Александр Андреевич. — Ведь прежде чем получить определение условной вероятности, мы вывели формулу (1.3) для случая равновозможных исходов, т. е. фактически доказали для этого случая теорему умножения вероятностей, а потом по определению обобщили ее на случай произвольных пространств элементарных исходов.

Давайте теперь напишем теорему умножения для трех событий, т. е. выведем формулу для $P\{ABC\}$.

Обозначим сначала событие BC через B'. Тогда по формуле (1.5) мы получим:

$$P\{ABC\} = P\{AB'\} = P\{B'\}\,P\{A/B'\} = P\{BC\}\,P\{A/BC\}.$$

По той же формуле можно получить

$$P\{BC\} = P\{C\}\,P\{B/C\},$$

поэтому

$$P\{ABC\} = P\{C\}\,P\{B/C\}\,P\{A/BC\}. \qquad (1.6)$$

Теперь вы легко напишите нужную формулу для произведения любого числа событий.

— Конечно, — кивнула Галя, — например,

$$P\{ABCD\} = P\{D\}\,P\{C/D\}\,P\{B/CD\}\,P\{A/BCD\}.$$

— Правильно, — подтвердил Александр Андреевич. — А сейчас мы получим одну из основных формул в теории вероятностей. Пусть события B_1, B_2,..., B_n образуют произвольную полную группу попарно несовместных, т. е. в сумме дают достоверное событие. (События B_k могут быть, например, всеми элементарными исходами.) Если теперь A — любое событие, то

$$AB_1 + AB_2 + \ldots + AB_n = A(B_1 + B_2 + \ldots + B_n) = AU = A,$$

и все эти события AB_1, AB_2, ... , AB_n попарно несовместны.

Так, например, $AB_1 \cdot AB_2 = AB_1B_2$, но $B_1B_2 = V$, поэтому

$$AB_1B_2 = AV = V.$$

Значит

$$P\{A\} = P\{AB_1\} + P\{AB_2\} + \ldots + P(AB_n\}.$$

Воспользовавшись в каждом слагаемом формулой (1.5), приходим к формуле полной вероятности

$$P\{A\} = P\{B_1\}P\{A/B_1\} + P\{B_2\}P\{A/B_2\} + \ldots$$
$$\ldots + P\{B_n\}\,P\{A/B_n\}. \qquad (1.7)$$

Эту формулу обычно применяют тэгда, когда условные вероятности $P\{A/B_k\}$ определить легче, чем вероятность $P\{A\}$.

Если события B_1, B_2,..., B_n интерпретировать как взаимоисключающие гипотезы, одна из которых обязательно верна, то формула (1.7) представляет собой разложение произвольного события A по этим гипотезам.

Поскольку

$$P\{B/A\} = \frac{P\{AB\}}{P\{A\}} = \frac{P\{B\}\;P\{A/B\}}{P\{A\}},$$

получаем формулу для любого $B_k (k = 1, 2, \ldots)$,

$$P\{B_k/A\} = \frac{P\{B_k\}P\{A/B_k\}}{P\{B_1\}P\{A/B_1\} + \ldots + P\{B_n\}P\{A/B_n\}}. \qquad (1.8)$$

Это так называемая формула Байеса для вероятностей гипотез. Она позволяет оценивать вероятность каждой из взаимоисключающих гипотез, если известен результат эксперимента (произошло событие A).

Давайте снова рассмотрим задачу с бракованными деталями.

Как вы помните, событие A заключается в том, что выбранная деталь бракованная. Гипотеза B_1 состоит в том, что выбранную деталь делал мастер, а B_2 — ученик (здесь $B_2=\overline{B_1}$).

Тогда нам известны $P\{B_1\}=\frac{100}{200}=\frac{1}{2}$; $P\{B_2\}=\frac{100}{200}=\frac{1}{2}$ и заданы $P\{A/B_1\}=\frac{5}{100}=\frac{1}{20}$ и $P\{A/B_2\}=\frac{1}{5}$.

Мы хотим определить $P\{B_1/A\}$. По формуле Байеса получаем

$$P\{B_1/A\}=\frac{P(B_1)\,P(A/B_1)}{P\{B_1\}\,P\{A/B_1\}+P\{B_2\}\,P\{A/B_2\}}=$$

$$=\frac{\frac{1}{2}\cdot\frac{1}{20}}{\frac{1}{2}\cdot\frac{1}{20}+\frac{1}{2}\cdot\frac{1}{5}}=\frac{5}{25}=\frac{1}{5}.$$

Если мы захотим узнать $P\{B_2/A\}$, то можем опять использовать формулу Байеса. Но можем воспользоваться и формулой (1.4):

$$P\{B_1+B_2/A\}=P\{B_1/A\}+P\{B_2/A\}-P\{B_1B_2/A\}.$$

Так как $B_1+B_2=U$, а $B_1B_2=V$, то

$$P\{(B_1+B_2)/A\}=P\{U/A\}=\frac{P\{AU\}}{P\{A\}}=\frac{P\{A\}}{P\{A\}}=1;$$

$$P\{B_1\times B_2/A\}=P\{V/A\}=\frac{P\{AV\}}{P\{A\}}=\frac{P\{V\}}{P\{A\}}=0$$

и $P\{B_2/A\}=1-P\{B_1/A\}=\frac{4}{5}$.

— Интересно, — спросил Стрелкин, — вот у нас было

$$P\{B_1\} = P\{B_2\} = \frac{1}{2},\ \text{а}\ P\{B_1/A\} = \frac{1}{5}$$

$$\text{и}\ P\{B_2/A\} = \frac{4}{5},$$

т. е. $P\{B_1/A\} < P\{B_1\}$, а $P\{B_2/A\} > P\{B_2\}$.

А разве может оказаться так, что условная вероятность будет равна безусловной? Насколько я помню, мы об этом не говорили.

— Может. Вернемся к примеру с ремонтными бригадами. Пусть событие A состоит в том, что электрики работают на втором электровозе, а B — механики работают на первом.

Тогда

$$P\{A\} = \frac{1}{3}\ \text{и}\ P\{A/B\} = \frac{P\{AB\}}{P\{B\}} = \frac{\frac{1}{9}}{\frac{1}{3}} = \frac{1}{3}.$$

— Да, но здесь особый случай, — возразил Стрелкин. — Ведь мы тогда говорили, что бригады посылаются независимо друг от друга, поэтому так и должно было получиться.

— Правильно, — сказал Александр Андреевич, — я специально взял этот случай. Вы сами пришли к выводу, что интуитивное представление о независимости двух событий должно выражаться в равенстве условной и безусловной вероятностей.

Действительно, равенство $P\{A/B\} = P\{A\}$ означает, что известие о том, что произошло событие B, на шансы появления события A не повлияло, т. е. событие A как бы не зависит от события B. Поэтому мы можем

дать точное математическое определение: событие A не зависит от события B, если

$$P\{A/B\} = P\{A\}. \tag{1.9}$$

— А почему вы говорите, событие A не зависит от события B? — спросил Стрелкин. — Разве может быть так, что событие A от B не зависит, а B от A зависит?

— Давайте это проверим, — сказал Александр Андреевич. — Пусть выполняется соотношение (1.9). Нам нужно обязательно проверить, верно ли, что в этом случае выполняется равенство

$$P\{B/A\} = P\{B\}. \tag{1.10}$$

Но

$$P\{B/A\} = \frac{P\{A\,B\}}{P\{A\}} = \frac{P\{B\}\,P\{A/B\}}{P\{A\}} = \frac{P\{B\}\,P\{A\}}{P\{A\}} = P\{B\}.$$

Значит, если A не зависит от B, то и B не зависит от A. Поэтому можно говорить просто о независимости событий A и B. Только надо помнить, что соотношение (1.9) имеет смысл, когда $P\{B\} \neq 0$, а (1.10) — когда $P\{A\} \neq 0$.

Из сказанного следует, что вероятность произведения двух независимых событий (в формулу (1.5) надо подставить вместо условной вероятности $P\{B/A\}$ равную ей величину $P\{A\}$) равна

$$P\{AB\} = P\{A\}\,P\{B\}.$$

— Интересно, — заметила Галя. — Выходит, независимые события не могут быть несовместными, ведь если

$P\{A\} > 0$ и $P\{B\} > 0$, то получается, что $P\{AB\} > 0$, а если бы A и B были несовместны, то было бы

$$P\{AB\} = 0.$$

— Стоило столько считать! — Борис снисходительно посмотрел на Галю. — И так ясно, что если события A и B несовместны и событие A произошло, событие B произойти не может и $P\{B/A\} = 0$. Если $P\{B\} > 0$, то, конечно же, справедливо неравенство

$$P\{B/A\} \neq P\{B\}.$$

— Правильно, — одобрительно сказал Александр Андреевич. — Нельзя путать независимость и несовместность событий. Несовместность вообще не зависит от того, как введены вероятности на пространстве элементарных исходов.

— А как же быть, если у нас есть три независимых события? — спросил Стрелкин. — Для них теорема умножения запишется так:

$$P\{ABC\} = P\{A\}\,P\{B\}\,P\{C\}. \qquad (1.11)$$

— Это не так ясно, как может показаться сразу, — возразил Александр Андреевич. — Во-первых, что вы имели в виду, когда говорили о трех независимых событиях?

— Как, что? — удивленно переспросил Стрелкин.— То же, что и раньше: событие A независимо от B и C и события B и C между собой независимы.

— Вот оно что! — улыбнулся Александр Андреевич. — Под независимостью событий A, B и C вы понимаете их попарную взаимную независимость. Так?

— Так, — согласился Стрелкин.

— Ну что ж, давайте немного посчитаем, — сказал Александр Андреевич, и все поняли, что сейчас произойдет что-то интересное. — Пусть на грузовую станцию прибыли четыре вагона. В одном из них находятся грузы, которые адресованы хозяйственному магазину, в другом — магазину сантехники, в третьем — складу запчастей для автомобилей, а в четвертом есть грузы, адресованные всем этим получателям.

Представим себе, что ночью на станцию пробрались грабители, которые, естественно, желают украсть наиболее дефицитные товары. Но так как они не знают, что в каком вагоне находится, они вскрывают первый попавшийся вагон, наудачу.

Рассмотрим события A, B и C, состоящие в том, что в ограбленном вагоне оказались грузы, предназначенные соответственно хозяйственному магазину, магазину сантехники и складу запчастей. Каковы их вероятности?

— Такие вероятности мы уже много раз считали, — быстро сказал Борис. — Всего возможных исходов 4— по числу вагонов. Каждому из этих событий благоприятны два — вагон с соответствующим товаром и смешанный. Значит, все эти события имеют одинаковые вероятности — по $\frac{1}{2}$:

$$P\{A\} = P\{B\} = P\{C\} = \frac{1}{2}.$$

— Очень хорошо, — ответил Александр Андреевич. — А теперь посчитаем вероятности событий AB, AC, BC. Ну-ка, Галя!

— Событие AB заключается в том, что в ограбленном вагоне были товары для хозяйственного магазина и для сантехники, — бодро начала Галя. — Это возможно только при элементарном исходе, если ограблен сме-

шанный вагон. Значит $P\{AB\}=\frac{1}{4}$. И точно так же $P\{AC\}=P\{BC\}=\frac{1}{4}$.

— Прекрасно! Так, зависимы события A, B и C попарно?

— Независимы, — ответил Стрелкин. — Ведь, например,

$$P\{A/C\}=\frac{P\{AC\}}{P\{C\}}=\frac{\frac{1}{4}}{\frac{1}{2}}=\frac{1}{2}=P\{A\}$$

и также для остальных пар (A, B) и (B, C).

— Значит, события A, B и C попарно независимы. Согласны? — подвел итог Александр Андреевич. — А теперь возьмемся за событие ABC, которое состоит в том, что в ограбленном вагоне оказались товары для всех адресатов. Это опять-таки возможно лишь в одном случае, если ограблен смешанный вагон.

И, значит,

$$P\{ABC\}=\frac{1}{4}.$$

Вы согласны?

— Да! — хором ответили остальные.

— Но ведь $P\{A\}\times P\{B\}\times P\{C\}=\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{8}\neq\frac{1}{4}$, — с улыбкой сказал Александр Андреевич.

— Как же так? — удивленно спросил Стрелкин. — Вот, видите, как я и говорил раньше, подгонять под ответ приходится.

— Что ж, давайте разбираться. Возьмем формулу (1.6), которая верна для любой тройки событий, и

заменим в ней $P\{B/C\}$ на $P\{B\}$, что одно и то же, если события B и C независимы. Получим

$$P\{ABC\} = P\{C\}\,P\{B\}\,P\{A/BC\}.$$

Чтобы вероятность произведения равнялась произведению вероятностей, нужно, чтобы

$$P\{A/BC\} = P\{A\},$$

т. е. чтобы события A и BC были независимы.

— А разве это не выполняется автоматически, если A независимо с каждым из событий B и C в отдельности? — спросил Стрелкин.

— Как мы увидели в разобранном примере, это не так. Поэтому вводят еще одно определение независимости многих событий в отличие от попарной независимости.

Три события A, B и C называются независимыми в совокупности, если, кроме попарной независимости, выполняется еще и соотношение (1.11).

Вообще события A_1, A_2,..., A_n называются независимыми в совокупности, если для любого набора k событий из них вероятность произведения равна произведению вероятностей:

$$P\{A_{i_1} A_{i_2} \ldots A_{i_k}\} = P\{A_{i_1}\} \ldots P\{A_{i_k}\},$$

$$k = 2, 3, \ldots, n.$$

Это требование включает в себя при $k=2$ и попарную независимость.

Таким образом, изучив достаточно подробно свойства вероятностей, мы можем подойти к вопросу о правильности задания вероятностей элементарных исходов.

Сначала вообразим, что мы проделали два эксперимента. В первом может произойти событие A, а во втором — B. И пусть нам известно, что вероятность появления A в первом эксперименте равна p_1, а вероятность появления B во втором — p_2.

Теперь рассмотрим новый эксперимент, состоящий в том, что оба эксперимента проводятся подряд, и результатом этого нового эксперимента является пара результатов составляющих экспериментов. Так, например, результатами являются такие события:

$$\{A, B\},\ \{A, \overline{B}\},\ \{\overline{A}, B\} \text{ и } \{\overline{A}, \overline{B}\}.$$

Легко понять, что все эти четыре события образуют полную группу и могут рассматриваться как элементарные исходы нового эксперимента. Какие вероятности им присвоить?

— Любые! — ответила Галя. — Только чтобы были положительными и в сумме давали единицу.

— Ну нет, — не согласился Борис. — Предположим, что мы присвоили вероятности так: $P\{A, B\} = p_{11}$, $P\{A, \overline{B}\} = p_{12}$. Тогда, очевидно, событие $\{A, B\} + \{A, \overline{B}\}$ состоит в том, что в первом эксперименте произошло событие A, т. е. вероятность этого события должна быть равна p_1, а так как $\{A, B\}$ и $\{A, \overline{B}\}$ несовместны, то вероятность их суммы равна $p_{11} + p_{12}$.

Значит, должно быть $p_{11} + p_{12} = p_1$. И точно так же, если

$$P\{\overline{A}, B\} = p_{21},\ P\{\overline{A}, \overline{B}\} = p_{22},$$

то

$$p_{11} + p_{21} = P\{A, B\} + P\{\overline{A}, B\} = P\{B\} = p_2;$$

$$p_{21} + p_{22} = P\{\overline{A}, B\} + P\{\overline{A}, \overline{B}\} = P\{\overline{A}\} = 1 - p_1;$$

$$p_{12} + p_{22} = P\{A, \overline{B}\} + P\{\overline{A}, \overline{B}\} = P\{\overline{B}\} = 1 - p_2,$$

т. е. для четырех неизвестных p_{11}, p_{12}, p_{21}, p_{22} имеем четыре уравнения, ведь p_1 и p_2 нам заданы. Значит, есть только один возможный способ задания этих вероятностей.

— Так ли это? — заметил Александр Андреевич. — Рассуждения, конечно, верные, но Боря слишком увлекся и в конце ошибся. Давайте еще раз выпишем эти четыре уравнения:

$$p_{11} + p_{12} = p_1;$$

$$p_{11} + p_{21} = p_2;$$

$$p_{21} + p_{22} = 1 - p_1;$$

$$p_{12} + p_{22} = 1 - p_2.$$

Если учесть, что $p_{11}+p_{12}+p_{21}+p_{22}=1$, то третье уравнение легко получается из первого, а четвертое — из второго.

Так что следует рассматривать только три основных уравнения:

$$p_{11} + p_{12} = p_1;$$

$$p_{11} + p_{21} = p_2;$$

$$p_{11} + p_{12} + p_{21} + p_{22} = 1.$$

Для того чтобы получить все четыре вероятности, достаточно задать одну — например, p_{11}.

Тогда

$$p_{12} = p_1 - p_{11};$$

$$p_{21} = p_2 - p_{11};$$

$$p_{22} = 1 - p_{11} - (p_1 - p_{11}) - (p_2 - p_{11}) =$$
$$= 1 + p_{11} - p_1 - p_2.$$

А вот p_{11} мы можем задать произвольно, лишь бы выполнялись ограничения:

$$0 \leqslant p_{11} \leqslant p_1; \qquad (1.12)$$

$$0 \leqslant p_{11} \leqslant p_2;$$

$$p_{11} \geqslant p_1 + p_2 - 1.$$

— Меня заинтересовала одна деталь, — вступил в разговор Стрелкин. — Когда мы говорим о событии $\{A, B\}$, почему нельзя просто написать AB? Ведь это событие состоит в том, что одновременно произошли события A и B. Да ведь и раньше уже мы писали формулу

$$A = \{A, B\} + \{A, \overline{B}\},$$

правда, она у нас уже была в виде

$$A = AB + A\overline{B}$$

и, кроме того, p_{11} есть просто вероятность события AB.

— Это не совсем так, — ответил Александр Андреевич. — Дело в том, что события A и B являются возможными результатами разных экспериментов и, строго говоря, запись, которую вы предлагаете,

$$A = \{A, B\} + \{A, \overline{B}\}$$

смысла не имеет. Чтобы придать ей смысл, рассмотрим наш составной эксперимент и через A_0 обозначим событие, состоящее в том, что в первом эксперименте произошло событие A, а во втором — что угодно; через B_0 обозначим соответственно событие, состоящее в том,

что в первом эксперименте произошло что угодно, а во втором — событие B, т. е.

$$A_0 = \{A, U\}, \quad B_0 = \{U, B\}.$$

Событие A_0 есть результат нового составного эксперимента, которое на самом деле есть событие A, наступившее в первом эксперименте. Мы как бы описали старые эксперименты с помощью нового. Тогда можно написать:

$$\{A, B\} = \{AU, UB\} = \{A, U\} \cdot \{U, B\} = A_0 B_0;$$

$$A_0 = \{A, U\} = \{A, B + \overline{B}\} = \{A, B\} + \{A, \overline{B}\};$$

$$p_1 = P\{A, U\}, \quad p_2 = P\{U, B\}, \quad p_{11} = P\{A, B\} = P\{A_0 B_0\}.$$

Вернемся к заданию вероятности p_{11}.

Давайте в новом пространстве элементарных исходов — в пространстве пар — вычислим $P\{A/B\}$ или, как будет правильнее, — $P\{A_0/B_0\}$. Получим

$$P\{A_0/B_0\} = \frac{P(A_0 B_0)}{P(B_0)} = \frac{p_{11}}{p_2}.$$

Если мы выберем p_{11} так, чтобы

$$\frac{p_{11}}{p_2} = p_1,$$

т. е.

$$p_{11} = p_1 p_2 \tag{1.13}$$

(условия (1.12) будут выполнены, так как $0 \leqslant p_1, p_2 \leqslant 1$). Это будет означать, что события A_0 и B_0 независимы, т. е. появление A в первом эксперименте и появление B во втором независимы. Можно проверить, что в этом случае

все остальные три пары событий $\{A, \bar{B}\}$, $\{\bar{A}, B\}$ и $\{\bar{A}, \bar{B}\}$ независимы, т. е. независимы результаты первого и второго экспериментов. Соответственно вероятности пар будут вычисляться как произведения вероятностей. Например,

$$P\{\bar{A}, B\} = p_{21} = p_2 - p_{11} = p_2 - p_1 p_2 = (1 - p_1) p_2 = \\ = P\{\bar{A}\} P\{B\}.$$

Аналогично можно поступать и в случае любого числа экспериментов.

Таким образом, мы можем говорить о независимых экспериментах.

Если один и тот же эксперимент производится дважды, то происходит повторное независимое испытание. Если элементарными исходами какого-либо эксперимента являются события $A_1, A_2, \ldots, A_n$, которые в одном испытании могут появиться с вероятностями $p_1, p_2, \ldots, p_n$ соответственно, то элементарными исходами эксперимента из N последовательных повторений являются всевозможные наборы $B_1, B_2, \ldots, B_N$, где каждое событие B_k есть одно из $A_1, A_2, \ldots, A_n$.

Поскольку на каждом из N мест может стоять любое из событий A_i ($B_k = A_i$ означает, что при k-м испытании произошло событие A_i), то всего различных наборов будет n^N.

Так как эти испытания были независимыми, следует вероятности элементарных событий — наборов задавать по формуле

$$P\{B_1, B_2, \ldots, B_N\} = P\{B_1\} \ldots P\{B_N\},$$

где

$$P\{B_k\} = P\{A_i\} = p_i, \text{ если } B_k = A_i.$$

Например, если мы бросаем игральный кубик три раза, естественно считать, что результат каждого бросания не зависит от результатов других бросаний. В каждом бросании имеются шесть равновозможных исходов (выпадание одной из 6 граней). Тогда в трех бросаниях имеется $6^3=216$ исходов и все они будут равновозможны, т. е. им необходимо приписать вероятность по $\frac{1}{6}\cdot\frac{1}{6}\cdot\frac{1}{6}=\frac{1}{216}$.

— Ну, хорошо, — сказал Борис. — Из такого определения вероятностей элементарных исходов в эксперименте из N испытаний следует, что любой элементарный исход, скажем, в первом испытании, независим с любым элементарным исходом во втором. Но, может быть, это не выполнено для составных событий, а ведь представление о независимости подразумевает независимость любых событий, происходящих в результате независимых испытаний?

— Верно, — ответил Александр Андреевич. — Это действительно требует проверки, но сделать ее несложно и вы можете попробовать провести необходимые математические преобразования сами.

Рассмотрим теперь эксперимент, в результате которого может произойти или не произойти некоторое событие A. Как определить его вероятность? Повторим этот эксперимент N раз и подсчитаем, сколько раз произошло событие A. Пусть это будет $k_N(A)$. Можно доказать, что величина $\frac{k_N(A)}{N}$, являющаяся частотой появления события A, будет приближаться к $P\{A\}$, если увеличивать число испытаний N. Точнее, вероятность того события, что разность $\left|\frac{k_N(A)}{N}-P\{A\}\right|$ будет отличаться от 0 на некоторую величину, стремится к 0 с ростом N, как бы мала эта величина ни была. Записывается это так:

для любого сколь угодно малого положительного числа ε

$$P\left\{\left|\frac{k_N(A)}{N} - P\{A\}\right| > \varepsilon\right\} \to 0 \text{ при } N \to \infty. \quad (1.14)$$

Это так называемый закон больших чисел для последовательности независимых испытаний.

Можно даже оценить скорость убывания этой вероятности

$$P\left\{\left|\frac{k_N(A)}{N} - P\{A\}\right| > \varepsilon\right\} < \frac{P\{A\}(1-P\{A\})}{N\varepsilon^2} \quad (1.15)$$

или, учитывая, что $0 \leqslant P\{A\} \leqslant 1$ и, следовательно,

$$P\{A\}(1-P\{A\}) \leqslant \frac{1}{4},$$

можно написать

$$P\left\{\left|\frac{k_N(A)}{N} - P\{A\}\right| > \varepsilon\right\} < \frac{1}{4N\varepsilon^2}. \quad (1.16)$$

Формула (1.15) носит название неравенства Чебышева (частный случай).

Теперь можно наметить путь для проверки правильности введенных вероятностей элементарных исходов. Пусть, например, мы присвоили событию A вероятность $p = \frac{1}{25}$.

Зададимся числом $\varepsilon > 0$, равным $\frac{1}{100}$, и сделаем, например, $N = 250\,000$ испытаний. Пусть при этом событие A появилось k раз. Тогда по формуле (1.16) у нас должно быть

$$P\left\{\left|\frac{k}{N} - p\right| > \frac{1}{100}\right\} < \frac{1}{4 \cdot \frac{1}{10\,000} \cdot 250\,000} = \frac{1}{100}.$$

Но

$$P\left\{\left|\frac{k}{N}-p\right|>\frac{1}{100}\right\}=1-P\left\{\left|\frac{k}{N}-p\right\}\leqslant\frac{1}{100}\right\},$$

так как события, заключающиеся в том, что k удовлетворяет неравенствам $\left|\frac{k}{N}-p\right|>\frac{1}{100}$ или $\left|\frac{k}{N}-p\right|\leqslant\frac{1}{100}$, взаимно дополнительны. Поэтому

$$P\{250\,000p-2500\leqslant k\leqslant 250\,000p+2500\}\geqslant 1-\frac{1}{100},$$

т. е. $P\{7500\leqslant k\leqslant 12\,500\}\geqslant\frac{99}{100}$.

Значит, если оказалось, что k выходит за пределы отрезка (7 500, 12 500), например, получилось $k=20\,000$, то это должно навести нас на мысль, что вероятность $p=\frac{1}{25}$ не совсем соответствует действительности, хотя, конечно, это могло произойти с вероятностью $\frac{1}{100}$.

— А почему мы не можем сразу так вычислять вероятность? — спросил Борис. — Проведем много раз эксперимент, подсчитаем $k_N(A)$ и поделим его на N.

— Можем, конечно, — ответил Александр Андреевич.— На практике часто так и поступают. Такое введение вероятности иногда называют статистическим или частотным, так как величина $\frac{k_N(A)}{N}$ выражает частоту появления события A в N испытаниях. Запишем неравенство Чебышева иначе:

$$P\left\{\frac{k_N(A)}{N}-\varepsilon\leqslant P\{A\}\leqslant\frac{k_N(A)}{N}+\varepsilon\right\}\geqslant 1-\frac{1}{4\varepsilon^2 N}, \quad (1.17)$$

т. е. если в результате N испытаний событие A произошло $k_N(A)$ раз, то мы с вероятностью по меньшей мере

$1 - \frac{1}{4\varepsilon^2 N}$ можем утверждать, что истинная вероятность $P\{A\}$ заключена в пределах $\left[\frac{k_N(A)}{N} - \varepsilon; \frac{k_N(A)}{N} + \varepsilon\right]$, эти пределы называют доверительным интервалом, а величину $1 - \frac{1}{4\varepsilon^2 N}$ — коэффициентом доверия.

— Все это хорошо, — иронически отозвался Стрелкин. — Но где взять время, чтобы проводить тысячи и миллионы испытаний? Кроме того, ведь есть эксперименты, которые вообще можно провести ограниченное число раз, а иногда нельзя провести вовсе. Если, например, мы интересуемся вероятностью разрушения оси вагона, не будем же мы разрушать тысячу вагонных осей для этого.

— Нет, конечно, — улыбнулся Александр Андреевич. — Здесь есть два ответа. Во-первых, закон больших чисел и неравенство Чебышева — довольно грубый математический инструмент. На них мы проследили лишь принципиальный подход к проблеме. Есть более точные теоремы. Об одной из них — центральной предельной теореме — мы еще будем говорить. Они позволяют при меньшем числе испытаний получать достаточно точные результаты. Кроме того, нет необходимости проводить эксперименты в натуре. Иногда их можно моделировать на вычислительных машинах, а здесь уже мы можем дать себе волю и повторять моделирование, сколько потребуется.

— Можно я вернусь немного назад? — спросил Борис. — К повторным испытаниям. Пусть, например, Галя бросает монету, у которой вероятность выпадания герба или решетки по $\frac{1}{2}$, и у нее так получилось, что 10 раз подряд выпал герб. Могло ведь так выйти?

— Конечно, — ответил Александр Андреевич. — Хотя вероятность этого события невелика.

— Вот-вот. Вероятность этого события должна быть $\left(\frac{1}{2}\right)^{10}$ — нужно десять раз перемножить вероятность события A. Ведь бросания монеты независимы. Если теперь Галя захочет бросить монету в одиннадцатый раз, то вероятность выпадания герба в этом бросании должна быть равна $\frac{1}{2}$. Верно?

— Верно, — подтвердил Александр Андреевич.

— Но это значит, что у нее одиннадцать раз подряд выпадет герб, а вероятность такого события равна $\left(\frac{1}{2}\right)^{11}=\frac{1}{2048}$.

— В чем же дело?

— А дело в том, что вы, Боря, забыли все, о чем мы говорили раньше. Вы опять перепутали условную и безусловную вероятности, — ответил Александр Андреевич. — $\left(\frac{1}{2}\right)^{11}$ — это вероятность того, что в одиннадцати независимых одинаковых испытаниях событие (выпал герб) произойдет одиннадцать раз. А если вы посчитаете вероятность того, что герб выпал в одиннадцатом испытании при условии, что он выпадал при каждом из первых десяти, то получите

$$P\{\text{герб в 11-м/герб при всех предыдущих 10}\} =$$

$$= \frac{P\{\text{герб при всех одиннадцати}\}}{P\{\text{герб при 10 первых}\}} = \frac{\left(\frac{1}{2}\right)^{11}}{\left(\frac{1}{2}\right)^{10}},$$

т. е. ту самую $\frac{1}{2} = P\{\text{герб в одном испытании}\}$, что полностью согласуется с независимостью испытаний.

Вернемся теперь к началу нашего разговора.

Когда Боря собирался идти в кассу за билетами, он мог предположить, что встретит очередь. Но заранее

точно сказать, сколько там будет человек, он, конечно, не мог. Это число зависит от многих причин, среди которых участвует и принятая им во внимание — то, что время отпусков еще не наступило. Но, видимо, другие причины оказали большее влияние. Число людей в очереди зависит от случая и принимает разные значения в зависимости от обстоятельств.

Точно так же зависит от многих, не подлежащих точному учету обстоятельств, число вызовов, поступающих на телефонный узел за определенный промежуток времени.

Нас не удивляет, что пуля при стрельбе попадает не точно в центр мишени — слишком много причин влияет на ее траекторию. Заранее предсказать отклонение от центра мы не можем — это тоже «случайная» величина. С такими величинами, которые под влиянием случайных обстоятельств способны принимать различные значения, также имеет дело теория вероятностей. Собственно случайные величины и являются самыми важными объектами исследования этой науки. Это и не удивительно. Ведь почти все величины, окружающие нас, в большей или меньшей степени являются случайными. Но чтобы все эти слова приобрели смысл, мы должны точно определить, что же это такое — «случайная» величина.

Для задания случайной величины нужно, во-первых, знать все значения, которые она может принимать. Число принимаемых случайной величиной значений может быть конечно или бесконечно. Эти значения могут быть разделены интервалами (число людей в очереди за билетами может принимать только целые значения) и заполнять целиком какой-нибудь промежуток, как, например, размеры рассеяния при стрельбе. Кроме того, нужно знать, с какими вероятностями случайная величина принимает свои значения.

— Как это так? — удивилась Галя. — Ведь вероятности мы приписывали событиям.

— Что же, мы и здесь будем делать то же самое,— ответил Александр Андреевич. — Представьте себе, что у Вас есть величина ξ, выражающая длину очереди за билетами. Она может принимать любое значение 0, 1, 2,...

— Ну не любое, — вставил Стрелкин. — Не может же, например, весь город стать в очередь.

— Да, конечно, но трудно ограничить точно длину очереди. Поэтому пока будем считать возможными любые значения. Теперь рассмотрим события E_0, E_1, ..., E_k, ... , состоящие в том, что длина очереди равна соответственно 0,1, ... , k, ... , т. е. $E_k = \{\xi = k\}$. Очевидно, что все эти события попарно несовместны и хотя бы одно из них произойдет. Значит, их можно принять за пространство элементарных исходов и вводить на нем вероятности: $p_k = P\{E_k\} = P\{\xi = k\}$.

— Понятно, — сказал Борис. — Теперь факт, что очередь не может быть бесконечной, выразится в том, что, начиная с некоторого номера N, все события E_{N+1}, E_{N+2}, ... будут иметь нулевую вероятность.

— Правильно, — подтвердил Александр Андреевич. — Возьмем теперь более сложный пример, скажем наш, с разрывом контактного провода. Примем за случайную величину η расстояние от начала участка M до точки разрыва L. Тогда эта величина может принимать значения от 0 до l, где l равно длине участка MM'. Как мы уже видели, здесь каждое значение принимается с вероятностью 0, но можно определить вероятность того, например, что $a \leqslant \eta \leqslant b$, где $0 \leqslant a < b \leqslant l$. При этом получаем $P\{a \leqslant \eta < b\} = \frac{b-a}{l}$.

А нам на самом деле вовсе и не требуется знать вероятность того, что некая величина, которая может

принимать все значения в некотором интервале, приняла точно определенное значение. Ведь всегда наши измерения связаны с погрешностью и достаточно знать вероятность того, что эта величина приняла значение, отличающееся от заданного не больше, чем на допустимую погрешность, т. е. $P\{c-\delta \leqslant \eta < c+\delta\}$. Значит, вероятности вида $P\{a \leqslant \eta < b\}$ нас вполне устраивают. Заметим также, что

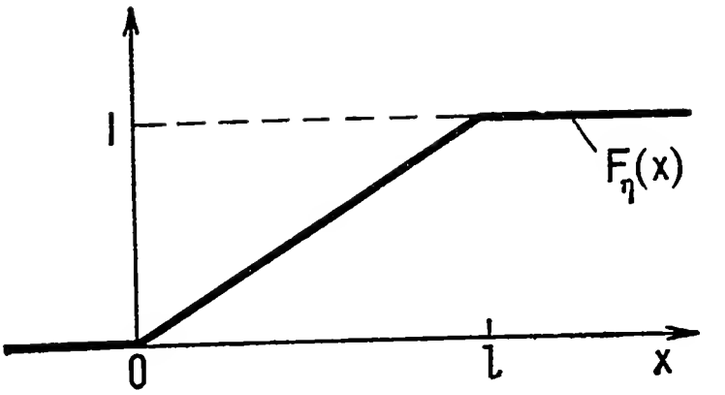


Рис. 1.3

$$\{-\infty < \eta < a\} + \{a \leqslant \eta < b\} = \{-\infty < \eta < b\}$$

и так как события в левой части этого равенства несовместны,

$$P\{-\infty < \eta < b\} = P\{-\infty < \eta < a\} + P\{a \leqslant \eta < b\}$$

или

$$P\{a \leqslant \eta \leqslant b\} = P\{-\infty < \eta < b\} - P\{-\infty < \eta < a\},$$

т. е. если мы будем при каждом x, заключенном между $-\infty$ и $+\infty$, знать величину

$$F_\eta(x) = P\{-\infty < \eta < x\},$$

то сможем вычислять и все требуемые нам вероятности. Для нашего примера получаем

$$F_\eta(x) = \begin{cases} 0 & \text{при } x \leqslant 0; \\ \dfrac{x}{l} & \text{при } 0 \leqslant x \leqslant l; \\ 1 & \text{при } x > l. \end{cases}$$

График этой функции изображен на рис. 1.3.

А теперь вернемся к примеру с очередью. Там мы тоже могли задавать не вероятности событий E_k, а функцию

$$F_\xi(x) = P\{\xi < x\}.$$

Событие $\{\xi < x\}$ состоит, очевидно, в том, что очередь не длиннее x, т. е. $\{\xi < x\} = E_0 + E_1 + \ldots + E_m$, где $m < x \leqslant m + 1$ и, значит,

$$P\{\xi < x\} = p_0 + p_1 + \ldots + p_m.$$

Тогда

$$P\{\xi = k\} = P\{\xi < k+1\} - P\{\xi < k\} = F_\xi(k+1) - F_\xi(k),$$

т. е. и в случае, когда случайная величина принимает дискретные значения, нам достаточно знать функцию $F(x)$. Только она будет принимать постоянные значения на отрезках между значениями 0, 1, 2,...

При этом длина очереди может оказаться и дробной, например, принимать значения между 1,3 и 1,7, но вероятность этого события равна:

$$P\{1{,}3 \leqslant \xi < 1{,}7\} = F_\xi(1{,}7) - F_\xi(1{,}3) =$$
$$= (p_0 + p_1) - (p_0 + p_1) = 0.$$

Итак, для задания случайной величины ξ нужно задать функцию $F_\xi(x) = P\{\xi < x\}$. Эта функция называется функцией распределения случайной величины ξ.

— Значит, случайная величина — это функция? — спросила Галя.

— Случайная величина — это величина, которая принимает некоторую совокупность значений с какими-

то вероятностями. Но задается случайная величина своей функцией распределения.

Давайте посмотрим, какими свойствами обладает функция распределения.

Во-первых, поскольку эта функция при каждом значении аргумента есть вероятность, то ее значения заключены между 0 и 1.

Во-вторых, эта функция не убывает, т. е. если $x_1 < x_2$, то

$$F_\xi(x_1) \leqslant F_\xi(x_2).$$

Действительно,

$$F_\xi(x_2) = P\{\xi < x_2\} = P\{\xi < x_1\} + P\{x_1 \leqslant \xi < x_2\} =$$

$$= F_\xi(x_1) + P\{x_1 \leqslant \xi < x_2\},$$

но так как $P\{x_1 \leqslant \xi < x_2\} \geqslant 0$, то $F_\xi(x_2) \geqslant F_\xi(x_1)$.

Если мы теперь рассмотрим событие $\{\xi < x\}$ и будем увеличивать x, то рано или поздно любое значение, которое ξ может принять, попадет в промежуток значений $(-\infty, x)$, т. е. событие $\{\xi < x\}$ при возрастании x будет приближаться к достоверному. Поэтому мы будем требовать, чтобы функция

$$F_\xi(x) \to 1 \text{ при } x \to +\infty.$$

По аналогичным причинам потребуем, чтобы $F_\xi(x) \to 0$ при $x \to -\infty$. Может оказаться, что при некотором x_1 $F_\xi(x_1) = 0$ или при некотором x_2 $F_\xi(x_2) = 1$.

Тогда из неубывания функции распределения следует, что при всех $x < x_1$ $F_\xi(x) = 0$ и при всех $x > x_2$ $F_\xi(x) = 1$.

— В примере с обрывом контактного провода у нас так и было, — напомнила Галя.

— Пусть теперь оказалось, что в точке x_0 функция $F_\xi(x)$ имеет разрыв (скачок), т. е. ее значения при приближении x к x_0 слева и справа стремятся к разным числам:

$$F_\xi(x_1) \to a \text{ при } x_1 \to x_0 \text{ слева}$$

$$\text{и } F_\xi(x_2) \to b \text{ при } x_2 \to x_0 \text{ справа.}$$

Поскольку при $x_1 < x_2$ $F_\xi(x_1) < F_\xi(x_2)$, то число a не может быть больше b и так как они не равны, то $a < b$. Что же это значит?

Рассмотрим разность $F_\xi(x_2) - F_\xi(x_1)$, где $x_1 < x_0 < x_2$. Эта разность равна вероятности события $\{x_1 \leqslant \xi < x_2\}$. При $x_1 \to x_0$ слева, а $x_2 \to x_0$ справа это событие стремится к событию $\{\xi = x_0\}$*.

Поэтому естественно считать, что

$$P\{x_1 \leqslant \xi < x_2\} \to P\{\xi = x_0\}$$

при $x_1 \to x_0$ слева
и $x_2 \to x_0$ справа

$$\text{Но } P\{x_1 \leqslant \xi < x_2\} = F_\xi(x_2) - F_\xi(x_1) \to b - a$$

при $x_1 \to x_0$ слева
и $x_2 \to x_0$ справа.

---

* Говорят, что последовательность событий $A_1 \supset A_2 \supset \ldots$, каждое из которых содержится в предыдущем, «сходится» к событию A, если $\bigcap_{n=1}^{\infty} A_n = A$, где $\bigcap_{n=1}^{\infty} A_n$ обозначает произведение всех событий A_n.

Значит, если функция $F_\xi(x)$ имеет в точке x_0 разрыв и скачок, равный p, то элементарный исход $\{\xi = x_0\}$ принимает не нулевую, а именно равную величине скачка p вероятность. Этот случай, если вы помните, как раз и встретился нам в примере с очередью за билетами.

Теперь надо решить еще один вопрос. Функция $F_\xi(x)$ должна быть задана при всех x. Какое же значение ей приписать в точке разрыва x_0? Ясно, что должно быть $a \leqslant F_\xi(x_0) \leqslant b$. Чтобы однозначно ответить на этот вопрос, рассмотрим события

$$\{\xi < x_1\},\ \{\xi < x_0\},\ \{\xi < x_2\},$$

где $x_1 < x_0 < x_2$.

Мы получим

$$F_\xi(x_0) - F_\xi(x_1) = P\{x_1 \leqslant \xi < x_0\};$$

$$F_\xi(x_2) - F_\xi(x_0) = P\{x_0 \leqslant \xi < x_2\}$$

при $x_2 \to x_0$ справа событие $\{x_0 \leqslant \xi < x_2\}$ стремится к событию $\{\xi = x_0\}$, а при $x_1 \to x_0$ слева событие $\{x_1 \leqslant \xi < x_0\}$ стремится к событию, не содержащему ни одного элементарного исхода $\{\xi = x\}$. Действительно, для любого $x < x_0$ обязательно найдется x_1 между x_0 и x, и событие $\{\xi = x\}$ окажется как бы вне события $\{x_1 \leqslant \xi < x_0\}$, а событие $\{\xi = x_0\}$ не входит ни в какое из событий $\{x_1 \leqslant \xi < x_0\}$,

$$F_\xi(x_2) - F_\xi(x_0) \to P\{\xi = x_0\} = b - a$$

и

$$F_\xi(x_0) - F_\xi(x_1) \to P\{V\} = 0,$$

а так как

$$F_\xi(x_2) \to b \text{ при } x_2 \to x_0 \text{ справа,}$$

то нужно определить $F_\xi(x_0) = a$.

Это свойство называется непрерывностью слева, а на графике обозначается стрелкой, как показано на рис. 1.4. Итак, всякая функция распределения обладает следующими свойствами:

1) $F_\xi(x)$ не убывает;
2) $F_\xi(x) \to 0$ при $x \to -\infty$;
   $F_\xi(x) \to 1$ при $x \to +\infty$;
3) $F_\xi(x)$ непрерывна слева.

Если в точке x_0 $F_\xi(x)$ разрывна (это значит, что

$$F_\xi(x) \to F_\xi(x_0 + 0) > F_\xi(x_0)$$

при $x \to x_0$ справа), то

$$F_\xi(x_0 + 0) - F_\xi(x_0) = P\{\xi = x_0\} > 0.$$

— А всякая ли такая функция есть функция распределения какой-нибудь случайной величины? — спросил Борис.

— Да, — ответил Александр Андреевич. — Для каждой функции с такими свойствами можно определить случайную величину так, чтобы ее функция распределения совпадала с данной функцией.

Итак, мы говорим, что ξ является случайной **величиной**, если можно вычислить вероятность каждого **из** событий $\{\xi < x\}$ при всех числовых значениях x.

— Можно, я перебью вас? — спросил Стрелкин. — Я хочу еще раз вернуться к примеру с очередью. Пусть

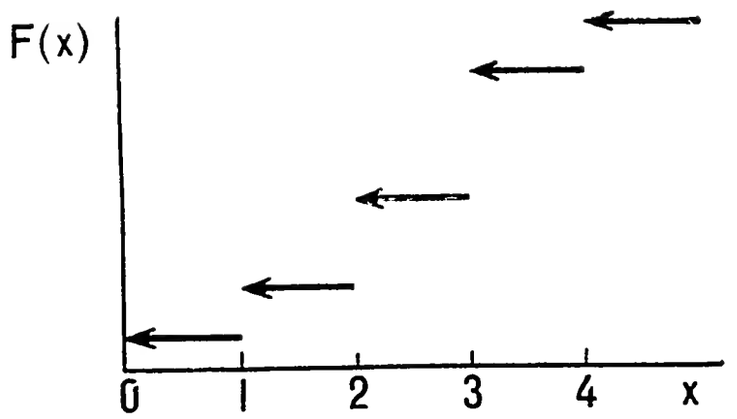


Рис. 1 4.

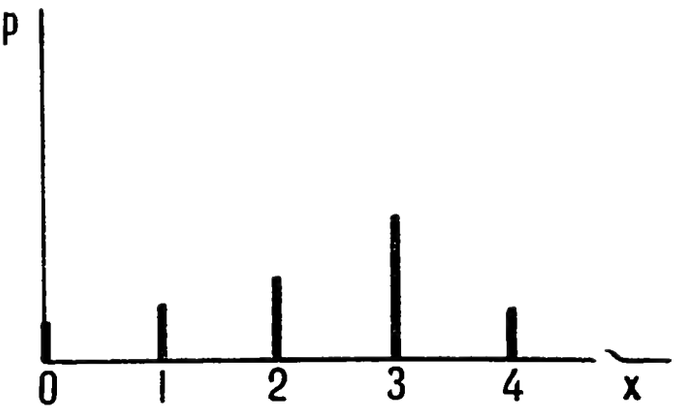


Рис. 1.5

случайная величина может принимать только значения 0, 1, 2, ..., как длина очереди. Тогда числа p_0, p_1 полностью определяют эту длину очереди как случайную величину. Правильно?

— Совершенно верно, — подтвердил Александр Андреевич. — Но и вероятности событий {длина очереди $<x$} также определяют эту величину.

— Это я понял, — продолжил Стрелкин. — Но числа как-то нагляднее, что ли. Вот скажем, две картинки (рис. 1.4 и 1.5). На первой нарисована функция распределения, а на второй только величины p_0, p_1,... именно в тех значениях x, которые наша величина принимает. Вторая картинка сразу подсказывает, что застать трех человек в очереди вероятнее, чем одного или двух. И хотя я понял, что функция распределения помогает описать одинаково и величины, которые принимают непрерывное множество значений и, как вы говорили, дискретные значения, мне кажется, что вторая картинка яснее.

— Вы правы, — ответил Александр Андреевич. — И мы сейчас попробуем получить столь же наглядную картинку в случае непрерывного множества значений случайной величины.

Давайте еще раз вернемся к вашему второму рисунку и попробуем разобраться, в чем его «нагляд-

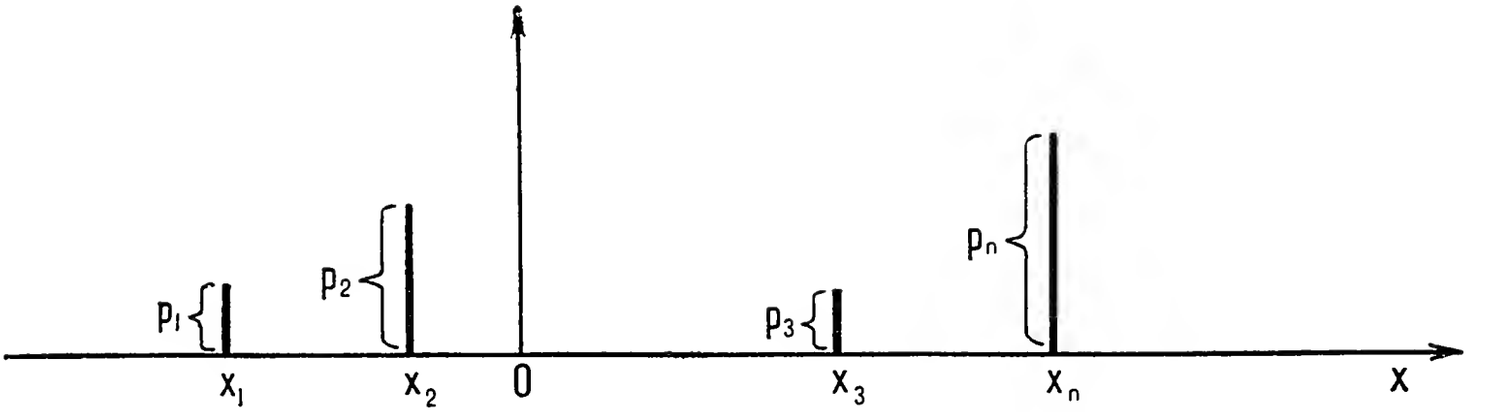


Рис. 1.6

ность». Но сначала перерисуем его для общей дискретной случайной величины. Пусть случайная величина ξ принимает значения $x_1, x_2, \ldots, x_n$, и только их с вероятностями соответственно $p_1, p_2, \ldots, p_n$. Тогда мы должны изобразить это таким образом, как показано на рис. 1.6.

Этот рисунок можно понимать как график некоторой функции от x, которая определена только при $x=(x_1, x_2, \ldots)$, причем значение этой функции при $x=x_n$ равно вероятности того, что случайная величина ξ принимает значение x_n, т. е. характеризует локальное свойство этой величины. Поскольку в случае непрерывного распределения вероятность каждого отдельного значения случайной величины равна нулю, чтобы построить такую же локальную характеристику, нам придется поступить несколько иначе.

Попробуем применить уже использованный однажды Борей прием. Пусть случайная величина ξ принимает непрерывное множество значений. Найдем характеристику этой величины при значениях, близких к числу x, т. е. рассмотрим вероятность того, что величина ξ приняла значения между x и $x+\Delta x$, где Δx — маленькое число. Мы уже знаем, что

$$P\{x \leqslant \xi < x + \Delta x\} = F_\xi(x + \Delta x) - F_\xi(x). \quad (1.18$$

— Теперь нужно поделить эту разность на Δx и перейти к пределу при $\Delta x \to 0$, — вмешался Борис. — Мы получим производную функции $F_\xi(x)$. Она и будет той функцией, которая нам нужна...

— Постойте, постойте, — остановил его Стрелкин. — Александр Андреевич, объясните, пожалуйста, это подробнее, если Борис прав.

— Да-да, сейчас объясню. Тем более, что все не так просто, как говорит Боря.

Наше равенство (1.18) можно понимать так, что на значения величины ξ между x и $x+\Delta x$ «приходится» вероятность $F_\xi(x+\Delta x) - F_\xi(x)$. Если Δx очень мало, то можно, очень грубо, конечно, себе представлять дело так, что на один исход приходится в среднем вероятность $\dfrac{F_\xi(x+\Delta x) - F_\xi(x)}{\Delta x}$. Но это нужно понимать не как вероятность одного элементарного исхода, которая равна нулю, а как «плотность» вероятности в точке x аналогично тому, как мы вводим понятие плотности вещества в точке, хотя масса точки равна нулю.

Таким образом, действительно имеет смысл рассматривать величину

$$p_\xi(x) = \lim_{\Delta x \to 0} \frac{F_\xi(x+\Delta x) - F_\xi(x)}{\Delta x}. \qquad (1.19)$$

Предел, который здесь написан, в математике называется производной функции $F_\xi(x)$ и обозначается через $F'_\xi(x)$. В этом Борис прав.

— Но разве всякая функция $F_\xi(x)$ имеет производную? — спросила Галя. — Ведь функция распределения может быть и разрывной, а мы знаем, что если у функции есть производная, то она обязательно непрерывна.

— Да, это очень важный момент, — ответил математик, — и здесь Боря уже проявил неосторожность. Действительно, если функция $F_\xi(x)$ разрывна, а это означает, как мы выяснили, что $F_\xi(x+0)=\lim\limits_{\Delta x\to 0} F_\xi(x+\Delta x)>F_\xi(x)$, то в правой части формулы (1.19) при $\Delta x>0$ в числителе будет величина, не меньшая, чем $F_\xi(x+0)-F_\xi(x)$, и так как знаменатель уменьшается до нуля, написанный предел будет равен ∞.

Следовательно, чтобы функция $P_\xi(x)$ была определена в точке x, необходимо, чтобы $F_\xi(x+0)-F_\xi(x)=0$. Значит величина ξ должна принимать значение x с вероятностью 0, так как $P\{\xi=x\}=F_\xi(x+0)-F_\xi(x)$.

Конечно, этого еще не достаточно. Могут быть и непрерывные функции, не имеющие производной, но сейчас мы не будем обсуждать этот вопрос, который относится к математическому анализу.

Пусть наша величина ξ такова, что ее функция распределения $F_\xi(x)$ имеет производную $p_\xi(x)=F'_\xi(x)$ при каждом x. Будем называть эту функцию $p_\xi(x)$ плотностью распределения и выясним, какой смысл имеет ее график.

Во-первых, так как $F_\xi(x)$ монотонно не убывает, то

$$F_\xi(x+\Delta x)-F_\xi(x)\begin{cases}\geqslant 0 \text{ при } \Delta x>0,\\ \leqslant 0 \text{ при } \Delta x<0.\end{cases}$$

Поэтому из (1.19) следует, что $p_\xi(x)\geqslant 0$. Возьмем теперь две близкие точки x и $x+\Delta x$ и будем считать, что между этими значениями функция $p_\xi(x)$ принимает

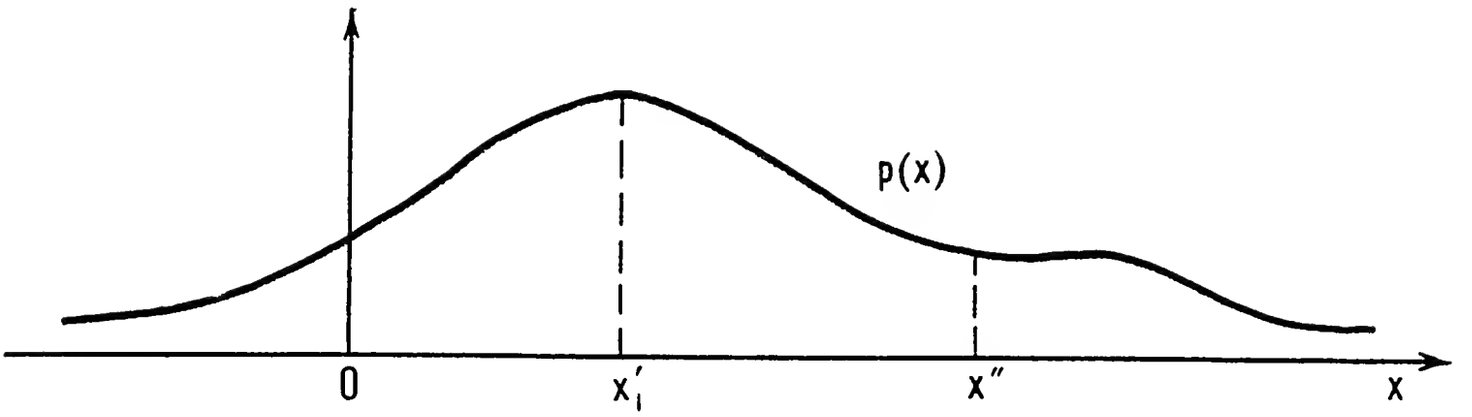


Рис. 1.7

одно и то же значение. Тогда площадь под кривой $p_\xi(x)$ между точками x и $x+\Delta x$ будет приблизительно равна

$$p_\xi(x)\,\Delta x \approx F_\xi(x+\Delta x) - F_\xi(x) = P\{x \leqslant \xi < x+\Delta x\}.$$

Это локальное свойство справедливо, оказывается, в любом промежутке. А именно, площадь S_{ab} под графиком плотности $p_\xi(x)$ численно равна вероятности $P\{a \leqslant \leqslant \xi < b\}$.

Для наших студентов я могу сказать, что это следует из геометрического смысла определенного интеграла. Ведь если

$$F'_\xi(x) = p_\xi(x) \text{ и } p_\xi(x) \text{ непрерывна,}$$

то

$$\int_a^b p_\xi(x)\,dx = F_\xi(b) - F_\xi(a) = P\{a \leqslant \xi \leqslant b\},$$

а интеграл, стоящий слева, численно равен площади S_{ab} под кривой $p_\xi(x)$.

— Значит, если я нарисую график плотности такой, как на рис. 1.7, то я смогу сказать, что величина ξ принимает значение x' с большей вероятностью, чем x''. Так? — спросил Стрелкин.

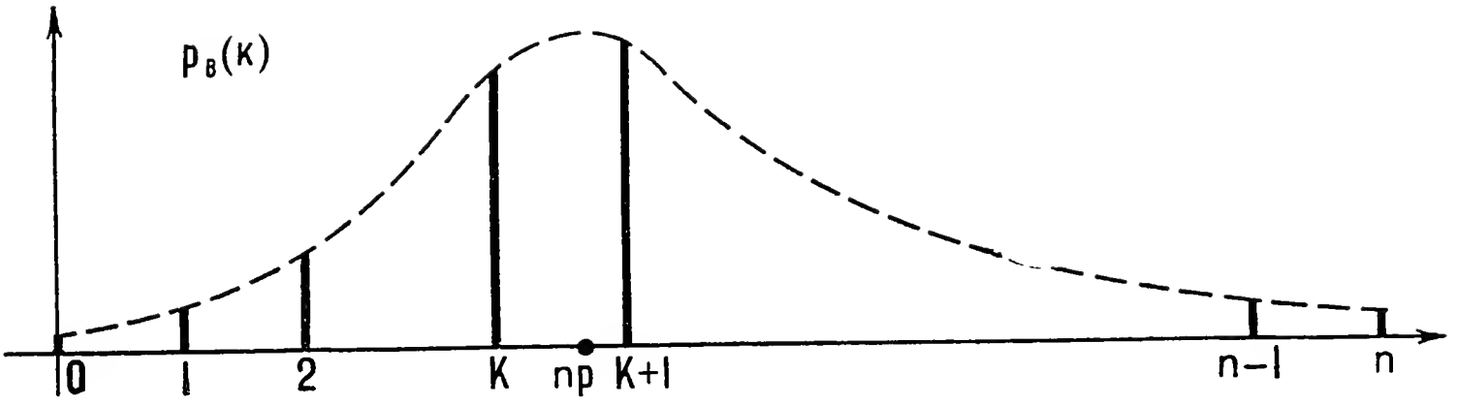


Рис. 1.8

— Почти так. Дело в том, что в точности значение x', как и значение x'', величина ξ принимает с вероятностью 0, но если вы скажете, что вероятность того, что ξ будет мало отличаться от x', больше, чем вероятность того, что ξ будет мало отличаться от x'', то будете абсолютно правы.

— Вы сказали, что плотность всегда положительна, — сказала Галя.

— Не положительна, а неотрицательна. Она может быть и нулем, — поправил Александр Андреевич.

— Ну да, но из того, что вы говорили дальше, видно, что площадь под всем графиком плотности должна быть равна 1. Так?

— Правильно, ведь если площадь под графиком между точками a и b равна $S_{ab}=P\{a\leqslant\xi<b\}$, то при $a\to-\infty$, $b\to+\infty$ мы получим, что вся площадь $S=P\{-\infty<\xi<\infty\}=1$. Эти два свойства — неотрицательность плотности и то, что площадь под ее графиком равна 1, являются определяющими. Любая функция, обладающая этими свойствами, может быть плотностью распределения случайной величины.

Итак, аналогом последовательности вероятностей значений дискретной случайной величины для непрерывной служит плотность распределения. Следовательно, всякая случайная величина полностью описывается

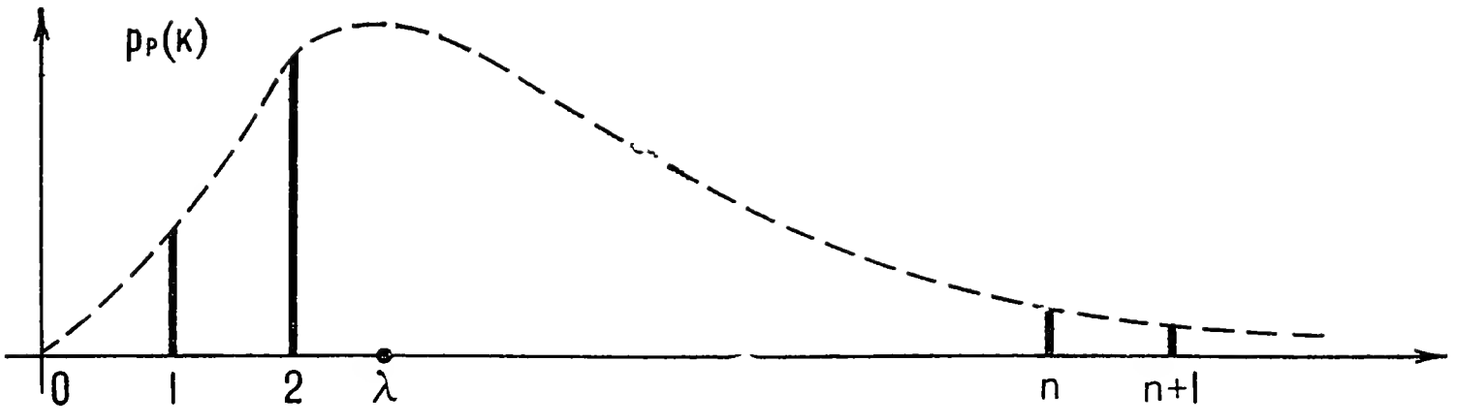


Рис. 1.9

законом распределения, т. е. любым соотношением, устанавливающим связь между возможными значениями случайной величины и соответствующими им вероятностями. Для дискретной случайной величины, как мы установили, такими законами являются: функция распределения, последовательность вероятностей ее значений или ряд распределения. Соответственно для непрерывной случайной величины — функция распределения и плотность распределения.

А сейчас, чтобы вы лучше поняли, я приведу несколько наиболее часто встречающихся в практике распределений. Думаю, что их будет полезно запомнить.

Дискретные распределения

1. *Биноминальное распределение* (параметры: натуральное число n и число p: $0 < p < 1$). Случайная величина принимает значения $0,1, \ldots, n$ и вероятность того, что она примет значение k, равна $p_B(k) = \frac{n!}{k!\,(n-k)!} \times \times p^k (1-p)^{n-k}$, где $n! = 1 \times 2 \times \ldots \times n$, $0! = 1$ по определению (рис. 1.8).

2. *Пуассоновское распределение* (параметр — число $\lambda > 0$). Случайная величина принимает значения 0, 1,

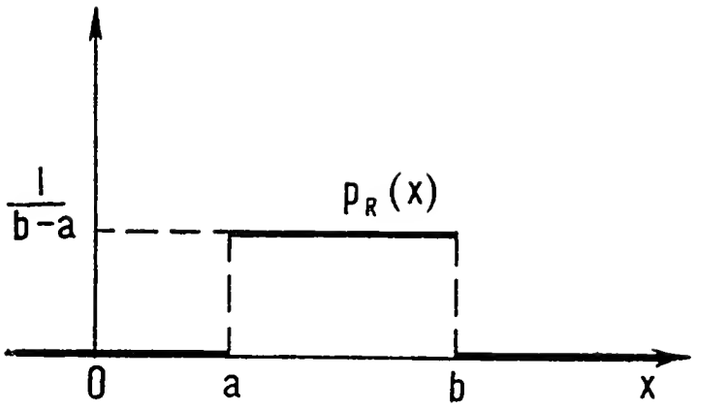


Рис. 1.10

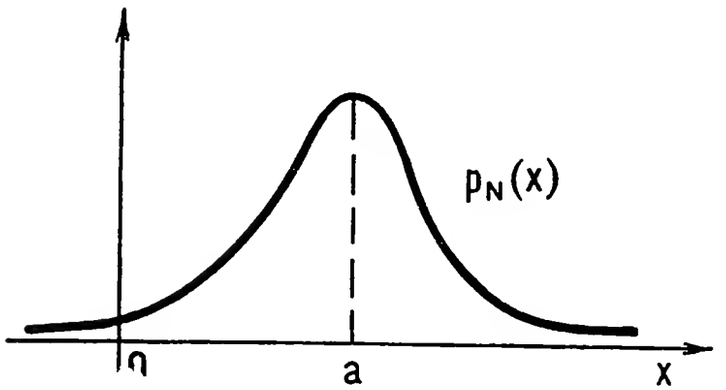


Рис. 1.11

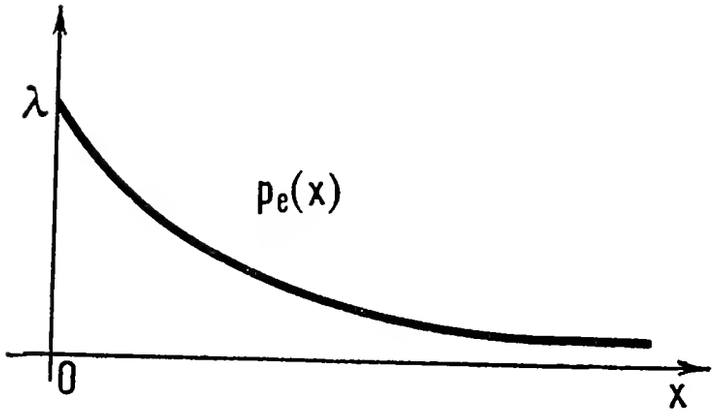


Рис. 1.12

2,..., и вероятность того, что она примет значение k, равна

$$p_P(k)=\frac{\lambda^k}{k!}e^{-\lambda},$$

где e — основание натуральных логарифмов, $e=2{,}718281828...$ (рис. 1.9).

Непрерывные распределения

3. *Равномерное распределение* на отрезке $[a, b]$.

Плотность задается функцией

$$p_R(x)=\begin{cases}\dfrac{1}{b-a}, & \text{если } a<x<b;\\ 0 & \text{в остальных случаях.}\end{cases}$$

График функции приведен на рис. 1.10.

4. *Нормальное распределение* или распределение Гаусса с числовыми параметрами a и $\sigma>0$ (рис. 1.11).

Плотность задается функцией

$$p_N(x)=\frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{(x-a)^2}{2\sigma^2}}.$$

5. *Экспоненциальное или показательное распределение (параметр* $\lambda > 0$).

Плотность задается функцией (рис. 1.12):

$$p_e(x) = \begin{cases} 0 & \text{при } x < 0 \\ \lambda e^{-\lambda x} & \text{при } x \geqslant 0. \end{cases}$$

— Ну хорошо, — сказал Стрелкин, — мы уже поняли, что случайную величину нужно характеризовать функцией распределения или набором вероятностей ее значений, или плотностью распределения. Но ведь в повседневной жизни мы то и дело говорим, что та или иная величина равна в среднем тому-то. Например, о той же очереди, которая есть величина случайная, мы говорим примерно так: «обычно в кассу очередь человек 10—15». Как это можно получить из всей той математики, о которой вы нам рассказали?

— Как раз самое время заняться различными характеристиками случайной величины. Обратимся к примеру. Что вы имеете в виду, говоря, что в среднем очередь составляет, скажем, 12 человек? — спросил Александр Андреевич Стрелкина.

— Это значит, что очередь может быть разной длины, и я как-то оценил возможность встретить эти различные случаи, например, я считаю, что у меня есть 20 шансов из 100 застать очередь в 8 человек, 20 шансов— 10 человек, по 30 шансов — 12 или 16 человек. Тогда я считаю, что в среднем я застану очередь

$$\frac{8 \cdot 20 + 10 \cdot 20 + 12 \cdot 30 + 16 \cdot 30}{100} = 12 \text{ человек.}$$

— Очень хорошо. Давайте теперь заменим ваши оценки вероятностями. Пусть случайная величина ξ, принимает дискретное множество значений $x_1, x_2, \ldots, x_n, \ldots$ с вероятностями $p_1, p_2, \ldots, p_n, \ldots$ соответственно.

Тогда средним значением этой случайной величины естественно называть сумму

$$x_1 p_1 + x_2 p_2 + \ldots + x_n p_n . \qquad (1.20)$$

Это число называется математическим ожиданием величины и обозначается $M\xi$.

— Да, — возмущенно сказала Галя, — но если, например, наша величина принимает значения 1, 2, 4, ... , 2^n, ... с вероятностями $\frac{1}{2}$, $\frac{1}{4}$, ... , $\frac{1}{2^{n+1}}$,..., то такая сумма $1\,\frac{1}{2} + 2\,\frac{1}{4} + \ldots + 2^n\,\frac{1}{2^{n+1}} + \ldots = \frac{1}{2} + \frac{1}{2} + \ldots + \frac{1}{2} + + \ldots$ равна ∞.

— Что же делать, — вздохнул Александр Андреевич. — Нам придется считаться с тем, что некоторые величины имеют бесконечно большие математические ожидания. Более того, некоторые величины могут и вовсе не иметь математического ожидания. Например, пусть случайная величина принимает значения -1, 2, -3, 4, -5, ... , $(-1^n)n$, ... с вероятностями p, $\frac{p}{2^2}$, $\frac{p}{3^2}$, ..., $\frac{p}{n^2}$, где $p > >0$, подобрано так, чтобы $p + \frac{p}{2^2} + \frac{p}{3^2} + \ldots + \frac{p}{n^2} + + \ldots = 1$.

Тогда, если мы будем вычислять сумму (1.20) так:

$$-1\cdot p + 2\cdot\frac{p}{2^2} - 3\cdot\frac{p}{3^2} + \ldots + (-1)^n n\cdot\frac{p}{n^2} + \ldots =$$

$$= -p + \frac{p}{2} - \frac{p}{3} + \ldots,$$

то мы получим число $-p\ln 2$, а если так:

$$-1\cdot p - 3\cdot\frac{p}{3^2} + 2\cdot\frac{p}{2^2} + 4\cdot\frac{p}{4^2} - 5\cdot\frac{p}{5^2} - 7\cdot\frac{p}{7^2} + 6\cdot\frac{p}{6^2} + \ldots,$$

то получим другое число. Какое же взять за математическое ожидание?

— Но, если, например, x_n и p_n такие, что

$$| x_1 | p_1 + | x_2 | p_2 + \ldots | x_n | p_n + \ldots < \infty, \qquad (1.21)$$

т. е. конечное число, то сумма (1.20) не зависит то того, в каком порядке брать слагаемые, — сказал Боря.

— Прекрасно! — воскликнул Александр Андреевич. — Действительно, с бесконечными суммами, или как их называют, рядами, нужно обходиться очень осторожно. В частности, нельзя, как в случае конечного числа слагаемых, менять их местами. Но если выполнено условие (1.21), то все операции переносятся и на случай бесконечного числа слагаемых. Поэтому мы будем говорить, что дискретная случайная величина ξ имеет математическое ожидание, равное сумме (1.20), если выполнено условие (1.21).

— А если случайная величина принимает только значения нуль и единицу, то чему тогда равно ее математическое ожидание? — спросила Галя.

— Тогда по формуле (1.20)

$$M(\xi) = x_1 p_1 + x_2 p_2 = 1 p_1 + 0 p_2 = p_1,$$

т. е. равно вероятности того, что случайная величина равна единице.

— Но нам известно уже, случайная величина может быть и непрерывной, — вставил Борис.

— В этом случае она имеет плотность $p_\xi(x)$, а ее математическим ожиданием согласно (1.20) будет выражение

$$M\xi = \int_{-\infty}^{\infty} x p_\xi(x)\, dx, \qquad (1.22)$$

если для плотности $p_\xi(x)$ выполнено условие, аналогичное (1.21)

$$\int_{-\infty}^{\infty} |x| p_\xi(x)\,dx < \infty. \qquad (1.23)$$

Теперь вспомним, что, говоря о длине очереди, наш уважаемый железнодорожник назвал не одно число, а некоторый интервал.

Верно?

— Да, — согласился Стрелкин, — я сказал, что средняя очередь не 12, а 10—15 человек, но я имел в виду, что на самом деле могут быть отклонения от среднего значения — 12 и наиболее вероятны отклонения не больше, чем на 3 человека, т. е. от 12—3 до 12+3, ну, вместо 9 я назвал 10.

— Что ж, это тоже характеристика случайной величины — среднее отклонение от среднего значения.

Как ее описать? Итак, у нас есть величина ξ и ее среднее значение $M\xi$. Тогда величину отклонения ξ от среднего значения можно записать как $|\xi - M\xi|$. Абсолютное значение берется потому, что ξ может быть и больше и меньше $M\xi$, а нас интересуют только размеры отклонения, а не его знак.

Поэтому среднее отклонение можно теперь записать как $M|\xi - M\xi|$.

В качестве характеристики отклонения величины ξ от $M\xi$ чаще используют так называемую дисперсию величины ξ, которая определяется следующим образом:

$$D\xi = M(\xi - M\xi)^2.$$

При этом само отклонение от среднего значения определяют величиной $\sqrt{D\xi}$, называемой средним квадратическим отклонением.

Отсюда видно, что дисперсия — всегда неотрицательная величина. Для дискретной случайной величины, принимающей значения $x_1, \ldots, x_n$ с вероятностями $p_1, \ldots, p_n$, дисперсия равна сумме

$$(x_1 - M\xi)^2 p_1 + (x_2 - M\xi)^2 p_2 + \ldots + (x_n - M\xi)^2 p_n + \ldots,$$

а для непрерывной величины с плотностью $p_\xi(x)$

$$D\xi = \int_{-\infty}^{\infty} (x - M\xi)^2 p_\xi(x)\, dx.$$

Для тех распределений, которые я выписывал, математическое ожидание и дисперсия выражаются через параметры следующим образом:

1) биномиальное распределение

$$M\xi_B = np, \quad D\xi_B = np(1-p);$$

2) пуассоновское распределение

$$M\xi_P = \lambda; \quad D\xi_P = \lambda;$$

3) равномерное распределение

$$M\xi_R = \frac{a+b}{2}, \quad D\xi_R = \frac{(b-a)^2}{12};$$

4) нормальное распределение

$$M\xi_N = a, \quad D\xi_N = \sigma^2;$$

5) экспоненциальное распределение

$$M\xi_e = \frac{1}{\lambda}, \quad D\xi_e = \frac{1}{\lambda^2}.$$

— Вы сказали, что дисперсия всегда неотрицательна, — заметил Стрелкин. — Но если она характеризует отклонения от среднего значения, то она, наверное, не может быть и нулевой. Ведь тогда это означало бы, что отклонения нет.

— Совершенно верно, — согласился математик. — Действительно, равенство $D\xi = M(\xi - M\xi)^2 = 0$ означает, что неотрицательная случайная величина $(\xi - M\xi)^2$ имеет среднее значение 0, а это возможно, только если она принимает положительные значения с вероятностью 0, а значение 0 с вероятностью 1, иначе сумма (1.20) или интеграл (1.22) были бы строго положительны. Значит $P\{(\xi - M\xi)^2 = 0\} = 1$, но событие $(\xi - M\xi)^2 = 0$ — то же самое, что и $\xi - M\xi = 0$, т. е. $\xi = M\xi$. Таким образом, если $D\xi = 0$, то $P\{\xi = M\xi\} = 1$, т. е. величина ξ с вероятностью 1 принимает только одно значение, равное $M\xi$. Такой случай называют вырожденным. Случайная величина, принимающая с вероятностью 1 одно значение, практически не является случайной.

Для невырожденной случайной величины дисперсия всегда строго положительна.

Пусть теперь у нас есть две случайные величины ξ и η с функциями распределения $F_\xi(x)$ и $F_\eta(x)$ соответственно.

Пусть произошло событие $\{\xi < x\}$. Что можно сказать о величине η?

— Из того, что нам пока задано, мы ничего сказать не можем, — ответил Борис.

— Но вот нам стала известна функция двух переменных

$$F_{\xi,\eta}(x, y) = P\{(\xi < x) \text{ и } (\eta < y)\},$$

т. е. заданы вероятности совместного наступления событий $(\xi < x)$ и $(\eta < y)$. Такую функцию $F_{\xi,\eta}(x, y)$ называют совместной функцией распределения. Если она нам

известна, то мы можем описывать взаимоотношение случайных величин ξ и η. Например, мы можем вычислить вероятность события $\{\eta < y\}$, если наступило событие $\{\xi < x\}$. Действительно, по определению условной вероятности

$$P\{\eta < y/\xi < x\} = \frac{P\{(\xi < x)(\eta < y)\}}{P\{\xi < x\}} = \frac{F_{\xi,\eta}\{x,\ y\}}{F_\xi(x)}.$$

— Значит, нам нужно знать не только $F_{\xi,\eta}(x, y)$, но и $F_\xi(x)$? — спросила Галя.

— Нет, $F_\xi(x)$ можно найти, зная $F_{\xi,\eta}(x, y)$, — возразил Боря. — Ведь $F_\xi(x) = P\{\xi < x\} = P\{(\xi < x)U\}$, где U — достоверное событие, например $U = \{\eta < \infty\}$. Поэтому $F_\xi(x) = P\{(\xi < x)(\eta < \infty)\} = F_{\xi,\eta}(x, \infty)$. Подставив в функцию $F_{\xi,\eta}(x, y)$ вместо y бесконечность, мы получим $F_\xi(x)$.

— Молодец, Боря, — похвалил Александр Андреевич, — только подставлять бесконечность в функцию не стоит. Просто нужно устремить y к $+\infty$, то есть потребовать выполнения равенства

$$F_\xi(x) = \lim_{y \to \infty} F_{\xi,\eta}(x, y).$$

Точно так же

$$F_\eta(y) = \lim_{x \to \infty} F_{\xi,\eta}(x, y).$$

Заметим теперь, что если события $\{\xi < x\}$ и $\{\eta < y\}$ независимы, то

$$F_{\xi,\eta}(x, y) = P\{(\xi < x)(\eta < y)\} = P\{\xi < x\}\,P\{\eta < y\} = F_\xi(x)\,F_\eta(y).$$

Если это равенство выполняется при всех x и y, т. е. если для любых x и y события $\{\xi < x\}$ и $\{\eta < y\}$ неза-

висимы, то случайные величины ξ и η называются независимыми.

Над случайными величинами можно совершать такие же действия, как и над обычными величинами — складывать, умножать, делить. При этом нужно уметь вычислять характеристики (в частности, математическое ожидание и дисперсию) результатов этих действий.

Оказывается, что математическое ожидание обладает следующими свойствами:

математическое ожидание произведения произвольного числа c и случайной величины ξ равно произведению этого числа на математическое ожидание случайной величины ξ

$$M(c\xi) = cM(\xi);$$

математическое ожидание суммы двух случайных величин ξ и η равно сумме математических ожиданий этих случайных величин

$$M(\xi + \eta) = M\xi + M\eta.$$

Но не всегда соблюдается равенство

$$M(\xi, \eta) = M(\xi) M\eta. \qquad (1.24)$$

Пусть, например, ξ и η одновременно принимают значения 1 с вероятностью $\frac{1}{2}$ и -1 с вероятностью $\frac{1}{2}$. Тогда $\xi\eta = 1$ с вероятностью 1 и $M(\xi\eta) = 1$, в то время как $M\xi = M\eta = 0$. А вот если величины ξ и η независимы, то равенство (1.24) справедливо. Правда, это равенство может выполняться и тогда, когда величины ξ и η зависимы. Пусть, например, ξ принимает значения 1 и -1, и

η — значения 0,1 и 2 с совместным распределением, определяемым вероятностями

$$P\{\xi=1,\ \eta=0\}=P\{\xi=1,\eta=1\}=P\{\xi=1,\ \eta=2\}=\frac{1}{6};$$

$$P\{\xi=-1,\ \eta=1\}=\frac{1}{2};$$

$$P\{\xi=-1,\ \eta=0\}=P\{\xi=-1,\ \eta=2\}=0.$$

Здесь ξ η принимает значения 1, 2, —1,0,—2 с вероятностями соответственно $\frac{1}{6}, \frac{1}{6}, \frac{1}{2}, \frac{1}{6}$,0 и, следовательно,

$$M(\xi\eta)=1\cdot\frac{1}{6}+2\cdot\frac{1}{6}+(-1)\cdot\frac{1}{2}+0\cdot\frac{1}{6}+(-2)\cdot 0=0.$$

А так как

$$P\{\xi=1\}=P\{\xi=1,\ \eta=0\}+P\{\xi=1,$$

$$\eta=1\}+P\{\xi=1,\ \eta=2\}=\frac{1}{6}+\frac{1}{6}+\frac{1}{6}=\frac{1}{2}$$

$$\text{и}\quad P\{\xi=-1\}=1-P\{\xi=1\}=\frac{1}{2},$$

то $M\xi=(-1)\cdot\frac{1}{2}+1+\frac{1}{2}=0$ и равенство (1.24) выполняется.

Но

$$P\{\xi<0,\ \eta<1\}=P\{\xi=-1,\ \eta=0\}=0,$$

а

$$P\{\xi<0\}=P\{\xi=-1\}=\frac{1}{2},\ P\{(\eta<1\}=P(\eta=0\}=\frac{1}{6}$$

$$\text{и}\quad P\{\xi<0\}\ P\{\eta<1\}=\frac{1}{12}\neq 0,$$

т. е. величины ξ и η не независимы.

Случайные величины, для которых справедливо равенство (1.24), называются некоррелированными. Теперь мы можем написать закон больших чисел в более общем виде, чем (1.14). Если $\xi_1, \xi_2, \ldots, \xi_n$ — последовательность попарно независимых одинаково распределенных случайных величин с математическим ожиданием M, то

$$P\left\{\left|\frac{\xi_1 + \ldots + \xi_n}{n} - M\right| < \varepsilon\right\} \to 1$$

при $n \to \infty$ для любого $\varepsilon > 0$, т. е. среднее арифметическое большого числа одинаково распределенных независимых случайных величин с большой вероятностью мало отличается от математического ожидания.

—Итак, — заключил Александр Андреевич, — мы познакомились с некоторыми характеристиками случайных величин. Следует лишь добавить, что не каждое из практически осуществимых событий «достойно» называться случайным. Закон больших чисел утверждает, что событие A, которое мы называем случайным и для которого определена вероятность, должно обладать некоторым свойством «устойчивости частот». Но не все события на практике обладают этим свойством. Некоторые события вообще могут не допускать даже мысленного повторения ситуации, в которой они могут проявиться. Но это уже вопросы, затрагивающие философскую сторону обоснования не только вероятностей, но и всей математики.

Я хотел бы еще отметить один замечательный факт, относящийся к случайным величинам.

Пусть у нас есть целая последовательность случайных величин $\xi_1, \xi_2, \ldots, \xi_n$, независимых и имеющих одинаковую функцию распределения;

m — их математическое ожидание, а σ^2 — дисперсия каждой из этих случайных величин.

Рассмотрим случайную величину

$$\eta_n = \frac{\xi_1 + \ldots + \xi_n - nm}{\sigma\sqrt{n}}.$$

Тогда

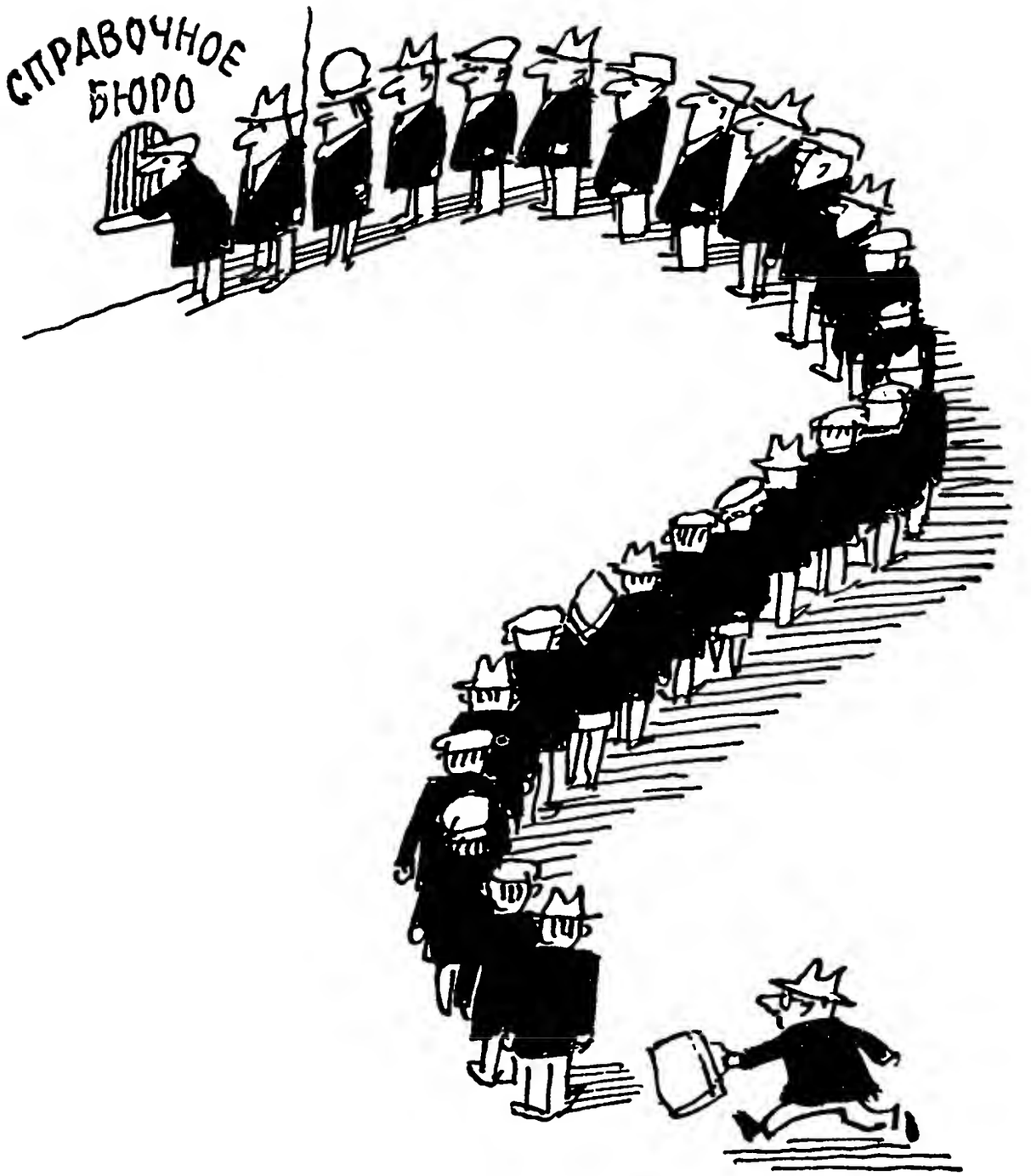
$$M\eta_n = \frac{1}{\sigma\sqrt{n}} M(\xi_1 + \ldots + \xi_n - nm) = 0;$$

$$D\eta_n = \frac{1}{\sigma_n^2} D(\xi_1 + \ldots + \xi_n - nm) = \frac{1}{\sigma_n^2}(D\xi_1 + \ldots + D\xi_n) = 1.$$

Если n достаточно велико, то распределение случайной величины η_n оказывается очень близко к нормальному распределению с математическим ожиданием равным 0 и дисперсией равной 1, т. е. имеющему плотность $\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

Этот факт носит название центральной предельной теоремы. Замечателен он тем, что распределение большого числа слагаемых не зависит от распределения каждого из них. Именно поэтому нормальное распределение так часто встречается в практических приложениях.

— Но я вижу, вы очень устали, — сказал Александр Андреевич, внимательно оглядев своих слушателей. — Сегодня мы, пожалуй, закончим. А завтра у нас новая тема.



БЕСЕДА ВТОРАЯ

Станем в очередь

(Теория массового обслуживания)

Утром Стрелкин пошел покупать газеты. Когда он добрался до киоска, там уже столпились пассажиры из соседних вагонов. Пришлось стать в очередь, хотя на успех рассчитывать было трудно — стоянка была короткой. Поезд тронулся, и Стрелкин по вагонам отправился к своему купе.

— Сегодня мы без газет, — объявил он соседям. — Была большая очередь.

— Ох, уж эти очереди, — вздохнула Галя, — в магазине — очередь, в кино — очередь, в парикмахерской — очередь. Как будто нельзя без них обойтись?

— К сожалению, нельзя, — сказал Александр Андреевич.

— Почему? Поставить в магазине пять касс вместо одной — вот и нет очереди.

— Да, но тогда у магазина могут быть большие убытки. Уменьшить очереди можно, кое-где их действительно можно ликвидировать, но сделать так, чтобы очередей вообще не было, нельзя. Заметьте, что стоят в очередях не только люди: стоят корабли перед разгрузкой, письма перед сортировкой, даже дома, которые нужно отремонтировать. Очереди встречаются гораздо чаще, чем вы можете представить, и в самых неожиданных местах.

— Неужели и этим занимается математика?

— Да. Есть даже раздел теории вероятностей, который так и называется — теория очередей, или теория массового обслуживания.

— Для чего? — удивился Стрелкин. — Ведь все очереди одинаковы, разве только одни короче, другие длиннее.

— Ну что вы! Существует много различных типов очередей. Если хотите, я могу назвать некоторые из них.

— Пожалуйста. Мы с удовольствием послушаем — тем более читать все равно нечего.

— Прежде всего, раз мы заговорили о теории массового обслуживания, позвольте мне воспользоваться некоторыми терминами. Первый из них — *требование* (или *заявка*). Требования — это элементы, из которых состоят очереди, — люди, пароходы, поезда, самолеты. Требования стоят в очереди в ожидании обслуживания. Что это такое, понятно — обслуживаются клиенты в парикмахерских, неисправные телевизоры в мастерских, поезда на станциях. Правда, в некоторых случаях применение этого термина может вызвать у вас улыбку: например, самолеты обслуживаются зенитками противника, банки обслуживаются бандами грабителей, которые в свою очередь обслуживаются полицейскими. Впрочем ситуация «полицейские и воры» относится скорее к теории игр. Обслуживание производится приборами (каналами). В наших примерах прибор — это парикмахер, радиомастер, сортировочная горка на станции, зенитка, банда воров.

Требования, поступающие в систему, образуют *входящий поток* требований. Требования, покидающие систему, — *выходящий поток*. Для систем, в которых некоторые требования по каким-то причинам не обслуживаются, различают выходящий поток обслуженных требований и выходящий поток необслуженных требований.

Входящий поток требований, очередь, обслуживающие приборы и выходящий поток составляют в совокупности систему массового обслуживания.

Теперь, когда мы вооружились некоторыми терминами, можно поговорить и о том, какие бывают очереди и системы.

Прежде всего, системы массового обслуживания можно разделить на *системы без очереди* и *системы с очередью* (рис. 2.1). Если в системе не бывает очереди, то возможны два варианта. Первый —

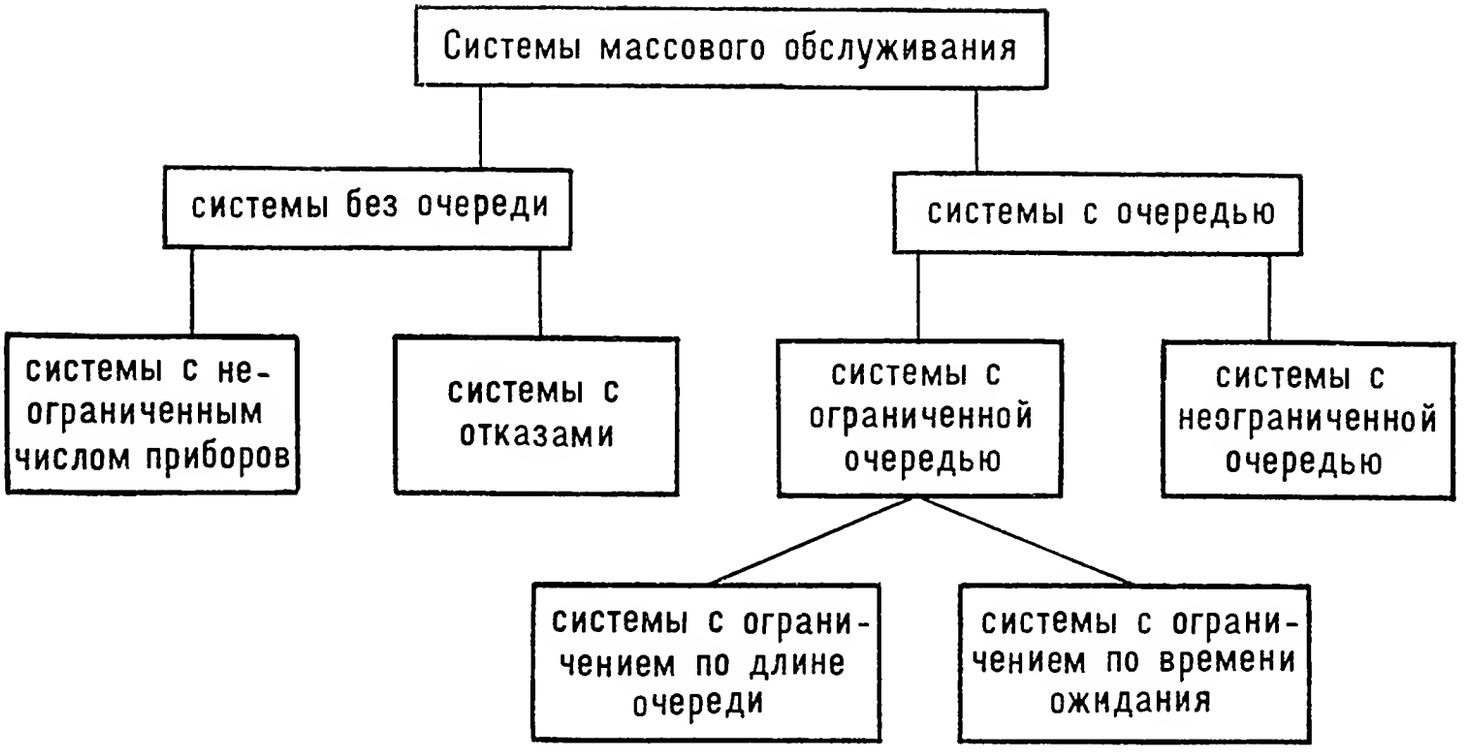


Рис. 2.1.

это система с отказами — требование, застав все приборы занятыми, покидает систему необслуженным. Когда вы, набрав первые несколько цифр нужного вам номера телефона, слышите короткие гудки, это значит, что на АТС все линии связи заняты. Вы получили отказ. Второй вариант — система без отказов — с неограниченным числом приборов. В жизни таких систем, конечно, не бывает, однако, когда число приборов намного больше возможного числа требований, удобно считать число приборов бесконечным. Примером может служить та же АТС в ночное время. Поток телефонных вызовов ночью невелик, и на АТС всегда найдется свободная линия для вашего разговора.

Системы с очередью, или с ожиданием, тоже можно разбить на две группы: системы с неограниченной очередью и системы с ограниченной очередью. В первом случае требование, попав в систему, не выйдет из нее необслуженным вне зависимости от того, в каком состоянии оно застало систему и сколько

ему придется ожидать в очереди. Например, мы с вами только что проехали мимо грузовой станции — все вагоны, которые мы видели, будут стоять в очереди, пока их не разгрузят. Во втором случае требование может уйти из системы необслуженным. Ограничения здесь бывают, как правило, либо по длине очереди (ограниченное число мест), либо по времени ожидания (очередь с нетерпеливыми клиентами). Примером системы с ограниченной длиной очереди может быть управляющая вычислительная машина, в которую поступают сообщения о состоянии некоторого объекта. Так как буферное запоминающее устройство имеет ограниченную емкость, в случае отсутствия в нем свободных ячеек вновь поступающие сообщения теряются. К очередям с нетерпеливыми клиентами относятся, по сути дела, все большие очереди, в которых нам приходится стоять. Вы, наверно, сами не раз уходили из очереди, устав ждать или когда спешили в другое место. Сегодня утром наш сосед попал именно в такую очередь, и, как вы знаете, оказался нетерпеливым клиентом.

— Будешь нетерпеливым, если не хочешь остаться в пижаме посреди чужого города, — отозвался Стрелкин.

— Надеюсь, вы теперь не скажете, что, все очереди одинаковы? Как видите, они существенно отличаются друг от друга, хотя то, о чем я рассказал, далеко не исчерпывает всего многообразия систем массового обслуживания. Есть и другие признаки, по которым их можно различать.

— Какие?

— Например, дисциплина обслуживания, т. е. порядок, в котором требования из очереди поступают на обслуживание. В тех очередях, с которыми мы ежедневно сталкиваемся, требования обычно обслуживаются в порядке поступления. Однако может быть и

другой порядок — например, инверсионный — требования обслуживаются в порядке, обратном порядку поступления, или случайный — требования выбираются из очереди случайным образом. Кроме того, важный класс составляют системы, в которые поступают требования с различными приоритетами. В таких системах требования с более высоким приоритетом обслуживаются в первую очередь. Так, на телеграфе первыми посылают срочные телеграммы, через перекресток раньше других пропускают пожарные и санитарные машины, самолеты, идущие на посадку, имеют преимущество перед самолетами, готовящимися взлетать.

Системы массового обслуживания различаются и по структуре обслуживающего устройства. Бывают системы одноканальные — с одним прибором и многоканальные — с несколькими приборами. О теоретическом случае бесконечного числа приборов мы с вами уже говорили. Обслуживание может быть однофазовым и многофазовым, если требование после обслуживания на первой фазе поступает на вторую, перед которой тоже может быть очередь, и так далее. Например, многофазовым является обслуживание в магазине — сначала вы выбиваете чек, потом получаете товар. Как правило, и там и там вам приходится стоять в очереди. Кстати, эта двухфазовая система часто бывает и многоканальной на одной или на обеих фазах: в магазине может быть несколько касс и (или) несколько продавцов, отпускающих один товар.

Несмотря на разнообразие систем обслуживания, любую, самую сложную из них, после достаточного изучения можно описать в терминах теории массового обслуживания и построить для нее вероятностную модель.

— Но что это дает? — спросил Борис. — Ведь такие системы надо уметь не только описывать, но и оптимизировать.

— Совершенно верно. И без теории массового обслуживания оптимизация во многих случаях невозможна. Что лучше — поставить вторую кассу в магазине самообслуживания или старую кассу заменить на новую, работающую в два раза быстрее? Результат, оказывается, совсем не один и тот же.

Какие величины могут интересовать нас при анализе систем массового обслуживания? Средняя длина и дисперсия очереди, среднее время пребывания в системе (или в очереди), среднее время занятости прибора. Для многоканальных систем, кроме того, среднее число одновременно работающих приборов и так далее.

— Смотрите, сортировочная горка, — указал на окно Стрелкин. — Интересно, Александр Андреевич, а сортировочную горку, наверно, можно считать системой массового обслуживания? Ведь она производит обслуживание — распускает составы.

— Прекрасно! — обрадовался Александр Андреевич, — мы делаем успехи. Если вы расскажете, как работает сортировочная горка, мы попытаемся представить ее как систему массового обслуживания.

— Ну что ж, попробую. Поезда, которые должны быть расформированы, поступают в парк прибытия сортировочной станции. Когда подходит очередь какого-то состава, специальный локомотив, работающий в этом парке, подает его на путь надвига и толкает состав до тех пор, пока все вагоны не разбегутся с горки по своим путям в парке формирования. Вот и все. Наверно, это очень простая система массового обслуживания.

— Не торопитесь. Иногда простые, на первый взгляд, системы оказываются не такими уж простыми. Что здесь является требованием, а что — обслуживающим прибором?

— Требования — это составы, а приборы — локомотивы. Локомотивов в парке прибытия может быть один

или два, поэтому система у нас получится одноканальной или двухканальной.

— Насчет требований вы правы, а вот относительно приборов я с вами не соглашусь. Ведь если бы локомотивов было не два, а десять, наверно, мало бы что изменилось, и система не стала бы десятиканальной?

— Да они только мешали бы друг другу!

— Мне кажется, — продолжал Александр Андреевич, — что число приборов определяется не числом локомотивов.

Сколько составов можно одновременно распускать на сортировочной горке?

— Обычно один.

— Следовательно, мы имеем дело с одноканальной системой массового обслуживания. Продолжительность обслуживания — это время роспуска состава. Эта величина, конечно, случайна, она зависит от длины состава, скорости движения толкающего локомотива.

— Александр Андреевич, — вмешалась Галя, — мне кажется, вы не совсем правы, когда говорите, что число локомотивов не имеет значения. Нас посылали на практику на сортировочную станцию и мы видели, как работает горка. Двух локомотивов вполне достаточно для того, чтобы, пока один локомотив распускает состав, второй локомотив мог вернуться с горки в парк прибытия и приготовился толкать следующий состав, как только предыдущий будет распущен. Добавление новых локомотивов работы не ускорит и в этом отношении нет разницы между двумя и десятью локомотивами. Но вот разница в работе горок с одним и двумя локомотивами есть, и немалая. — У меня есть предложение, — сказал Борис. — Надо считать фазу обслуживания двухэтапной: первый этап — это надвиг и роспуск состава, второй — возвращение локомотива в парк прибытия.

— Превосходно, молодые люди! — воскликнул Александр Андреевич. — Вы, несомненно, правы. Итак, мы считаем горку однофазной одноканальной системой массового обслуживания.

— Кажется, мы чего-то не учли, и наша модель не соответствует действительности, — сказал Борис. — Я несколько раз наблюдал такую картину: путь надвига свободен, горочные локомотивы тоже свободны, в парке имеются составы, а роспуск не начинается. Происходит задержка и иногда весьма значительная. А в нашей модели такой задержки не предусмотрено. В чем тут дело?

— Это я вам могу объяснить, — сказал Стрелкин.— Мы забыли про технический осмотр. Каждый поезд после поступления в парк прибытия обследуется бригадой технического осмотра, которая проверяет, все ли вагоны можно спускать с горки.

— А сколько времени продолжается этот осмотр? — поинтересовался Александр Андреевич.

— Это зависит от длины состава и технического состояния вагонов. Среднее время осмотра приблизительно равно или немного больше среднего времени, приходящегося на расформирование одного состава на горке.

— Сколько же таких бригад работает в парке прибытия?

— Одна или две, в зависимости от интенсивности поступления поездов на станцию. Каждый поезд осматривает только одна бригада. Обычно бригада старается закончить осмотр состава к моменту, когда придет пора распускать его с горки, но бывает так, что этого сделать не удается, и тогда случается, что горка свободна, а состав распускать нельзя.

— Это бывает, видимо, и тогда, — сказала Галя, — когда поезд приходит на станцию, а в это время

горка свободна и парк прибытия пуст. После появления поезда горка будет оставаться свободной до тех пор, пока не закончат осмотр прибывшего поезда.

— Обратите внимание на то, — сказал Александр Андреевич, — что технический осмотр — это первая фаза обслуживания, которая предшествует второй фазе обслуживания — роспуску состава.

— Странно, — удивилась Галя, — ведь составы, как будто, стоят в очереди на горку, а не на технический осмотр, стоит ли считать его отдельной фазой?

— Вы, Галя, допускаете ошибку. Да, действительно поезда стоят в парке прибытия для того, чтобы их расформировали. Это — главная цель. Но технический осмотр — обязательное звено технологии обработки составов и, следовательно, это такая же полноправная фаза обслуживания, как и сам роспуск. С точки зрения теории массового обслуживания, требования — составы — вначале стоят в очереди к первой фазе обслуживания — техническому осмотру (эта фаза одноканальная или двухканальная), а после обслуживания на ней переходят в очередь ко второй фазе — роспуску (эта фаза одноканальная). В реальной системе состав после технического осмотра никуда не переходит, а остается на том же пути до момента роспуска. Все это и ввело вас, Галя, в заблуждение. Но теперь, кажется, механизм обслуживания нам ясен. Обратим внимание на очереди — их две: перед первой фазой и перед второй. Давайте сначала выясним характер очереди перед первой фазой — техническим осмотром: она ограничена или нет?

— Конечно, ограничена, — сказал Борис. — Ограничена по длине: количество требований не может превышать числа путей в парке прибытия. Это ясно.

— Не совсем. Вспомните: в системах с очередью, ограниченной по длине, требования, заставшие все

места в очереди занятыми, для системы теряются и обслужены не будут. Что же бывает, когда поезд, предназначенный в расформирование на нашей станции, прибывает и застает все пути занятыми? Если наша система имеет ограниченную по длине очередь перед первой фазой, поезд будет для нашей станции потерян. Он что — направляется на другую сортировочную станцию?

— Нет! — вмешался Стрелкин. — Этого никогда не бывает. Если поезд должен быть расформирован на этой станции, то он на ней и будет расформирован.

— Так что же с ним происходит?

— Он ждет, но только на подходе — на предыдущей станции, например. Как только в парке прибытия освободится путь, его сразу примут туда.

— Ага, теперь понятно. Вы, Боря, тоже смешали реальную систему с математической моделью. На станции, в очереди к горке, можно увидеть, самое большое, столько поездов, сколько путей в парке прибытия; и вы решили, что очередь ограничена. На самом деле, очередь не ограничена, а места для ожидания — это парк прибытия плюс все возможные для ожидания пути на соседних станциях. Теперь скажите, какой будет очередь ко второй фазе?

— Ну уж тут-то очередь будет ограничена числом путей в парке прибытия, — ответил Борис.

— На этот раз вы правы. Составов, прошедших первую фазу — технический осмотр, но не попавших на горку, может быть максимум столько, сколько путей в парке прибытия.

— А что будет происходить с первой фазой — осмотром — если произойдет как раз такой случай? — спросила Галя.

— Эта фаза будет простаивать, хотя перед ней может быть очередь на соседних станциях, — ответил Стрелкин.

— Простой первой фазы при переполнении очереди ко второй фазе называется блокировкой, — заметил Александр Андреевич.

— Итак, друзья, подведем итоги, — Александр Андреевич оглядел попутчиков. — Простая, на первый взгляд, задача оказалась весьма непростой. Имеются две фазы обслуживания. Первая фаза одно- или двухканальная, вторая одноканальная. Очередь перед первой фазой не ограничена, перед второй ограничена. При переполнении второй очереди первая фаза блокируется.

— А теперь, — продолжал Александр Андреевич, — если вы заинтересовались, давайте возьмем бумагу и рассмотрим некоторые вещи подробнее. Первое, что нужно знать в массовом обслуживании, что такое входящий поток требований.

Интервалы, через которые требования поступают в систему, — случайные величины (в частном случае они могут быть постоянны: требования поступают через равные интервалы, образуя регулярный поток). Различные виды распределений длин этих интервалов и количества требований, одновременно поступающих в систему, порождают различные входящие потоки. Мы остановимся на некоторых потоках, которые встречаются наиболее часто. Для этого дадим сначала некоторые определения.

Поток требований называется стационарным, если вероятность поступления определенного количества требований за некоторый промежуток времени зависит только от длины этого промежутка (и не зависит от того, где на временной оси мы выберем этот промежуток). В природе стационарных потоков не существует, но очень многие из них могут без особой погрешности считаться стационарными или имеют стационарные (опять-таки с известной долей приближения) уча-

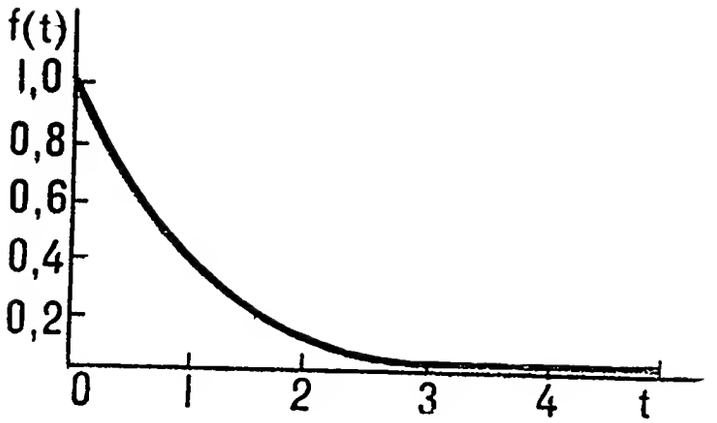


Рис. 2.2

стки. Поток пассажиров, желающих приобрести билет на поезд дальнего следования, днем и ночью не один и тот же (кассы работают круглосуточно). Однако между 10 и 14 часами этот поток, как правило, можно считать стационарным. То же самое можно сказать и о потоке телефонных вызовов. Стационарный поток имеет постоянную интенсивность поступлений. (Интенсивность — это математическое ожидание числа требований в единицу времени.) Мы будем говорить дальше только о стационарных потоках: во-первых, они очень часто встречаются (с той же оговоркой), а во-вторых, потому, что математический аппарат для систем с такими потоками разработан наиболее подробно.

Поток требований называется ординарным, если в любой момент времени может поступить только одно требование. Поток самолетов, взлетающих с аэродрома, ординарен, а вот поток взлетающих пассажиров неординарен, так как одновременно в самолете взлетает несколько пассажиров. Если в неординарном потоке одновременно поступает одно и то же количество требований, то поток очень легко свести к ординарному. Например, на грузовой пункт каждый раз поступает фиксированное число вагонов — 10. Если мы будем считать требованием не вагон, а подачу, поток превратится в ординарный.

Поток требований называется потоком с ограниченным последействием, если интервалы между поступлениями требований составляют последовательность взаимно независимых случайных величин.

Это означает, что если какой-нибудь интервал оказался особенно длинным или коротким, то это не скажется на длине других интервалов. Если такой поток стационарен и ординарен, то мы будем называть его рекуррентным потоком или потоком Пальма.

Первый рекуррентный поток, который мы рассмотрим, простейший или стационарный пуассоновский поток. Функция распределения интервалов между требованиями этого потока:

$$F(t) = 1 - e^{-\lambda t}, \tag{2.1}$$

λ называется параметром потока.

Плотность распределения интервалов (рис. 2.2) получается дифференцированием $F(t)$:

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{при } t \geqslant 0 \\ 0 & \text{при } t < 0. \end{cases} \tag{2.2}$$

Это уже известный нам показательный или экспоненциальный закон распределения.

Простейший поток занимает среди потоков такое же место, какое нормальный закон среди распределений. Так же, как при суммировании большого числа независимых случайных величин, получается распределение, близкое к нормальному, так и при наложении большого числа стационарных ординарных потоков получается поток, близкий к простейшему.

— А сколько потоков должно наложиться, чтобы получился простейший поток? — поинтересовался Борис.

— Теоретически число суммируемых потоков должно быть бесконечным, при этом интенсивность каждого потока должна стремиться к нулю. Практически поток, получаемый суммированием пяти-шести независимых потоков с соизмеримыми интенсивностями, уже может

считаться простейшим. Чем больше потоков налагается, тем лучше приближение суммарного потока к простейшему. Потоки автомобилей, подъезжающих к грузовой станции, потоки пассажиров, входящих в метро, потоки вызовов на телефонной станции близки к простейшим, так как составляющие их требования поступают из большого числа разных источников.

Кстати, именно развитие теории телефонной связи и явилось толчком для создания теории массового обслуживания. Первые исследования в этой области принадлежат датскому математику Эрлангу и относятся к 1909 г. Простейший поток исторически оказался, таким образом, первым исследованным потоком.

Простейший поток удобен тем, что для систем массового обслуживания с таким входящим потоком получаются наиболее простые решения. Хотя это может показаться странным, получить решение для системы, например, с регулярным входящим потоком труднее, чем для системы с простейшим потоком.

Простейший поток обладает одним замечательным свойством, которое называется отсутствием последействия. Оно заключается в том, что время, оставшееся до момента поступления нового требования, не зависит от того, сколько времени прошло после поступления последнего требования. Пусть, например, в систему поступает простейший поток со средним интервалом между требованиями 10 мин. После прихода очередного требования прошло 9 мин. Как вы думаете, Галя, через сколько минут (в среднем) придет новое требование?

— Наверно, через одну минуту?

— Нет! Средняя продолжительность оставшегося времени ожидания — 10 мин. Ну, а если случилось так, что с момента прихода последнего требования прошел час — когда придет новое требование?

— Я уже знаю, как надо ответить, — сказал Борис. — Среднее оставшееся время — опять-таки 10 мин. Но ведь это парадоксально!

— Тем не менее, это странное свойство можно очень легко доказать. Действительно,— Александр Андреевич взял карандаш, — пусть после прихода последнего требования прошло время a. Интервал между приходами последнего и ожидаемого требований — случайная величина ξ с функцией распределения (2.1). Нужно найти $P\{\xi - a < t \mid \xi \geqslant a\}$ — вероятность того, что оставшееся до прихода новой заявки время окажется меньше t при условии, что весь интервал больше a. По формуле условной вероятности, которая у нас уже была:

$$P\{\xi - a < t \mid \xi \geqslant a\} = \frac{P\{\xi - a < t \text{ и } \xi \geqslant a\}}{P\{\xi \geqslant a\}} = \frac{P\{a \leqslant \xi < t + a\}}{P\{\xi \geqslant a\}} =$$

$$= \frac{P\{\xi < t + a\} - P\{\xi < a\}}{P\{\xi \geqslant a\}} = \frac{1 - e^{-\lambda(t+a)} - (1 - e^{-\lambda a})}{e^{-\lambda a}} =$$

$$= \frac{e^{-\lambda a}(1 - e^{-\lambda t})}{e^{-\lambda a}} = 1 - e^{-\lambda t} = F(t), \qquad (2.3)$$

так как

$$P\{\xi \geqslant a\} = 1 - P\{\xi < a\} = e^{-\lambda a}, \qquad (2.4)$$

т. е. условная функция распределения времени, оставшегося до прихода новой заявки, равна функции распределения целого интервала между поступлениями требований.

Из рекуррентных потоков (стационарных потоков с ограниченным последействием) отсутствие последействия свойственно только простейшему потоку. Остальные потоки имеют ограниченное последействие. Эта ограниченность выражается в независимости длины интервала

между поступлениями требований от длины предыдущих интервалов.

— А почему простейший поток по-другому называется стационарным пуассоновским потоком? — спросила Галя.

— Дело вот в чем. Мы задали простейший поток функцией распределения интервалов между поступлениями требований. Существует другой способ определения простейшего потока с помощью распределения вероятностей поступления определенного количества требований в заданном промежутке времени. Оба эти способа равноправны, из одного следует другой.

Действительно, по индукции можно доказать, что если функция распределения интервалов между поступлениями требований в потоке имеет вид (2.1), то число требований, поступивших за промежуток времени t, распределено по закону Пуассона (поэтому простейший поток называется по-другому стационарным пуассоновским):

$$P_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}, \qquad (2.5)$$

где $P_k(t)$ — вероятность того, что за промежуток времени t поступит ровно k требований. Для всевозможных значений k получается семейство функций $P_k(t)$ ($k = 0,1, 2, \ldots$).

Если простейший поток задан вторым способом, то нетрудно получить функцию распределения интервалов между поступлениями заявок. Пусть случайная величина ξ — длина интервалов между требованиями. Тогда

$$F(t) = P\{\xi < t\} = 1 - P\{\xi \geqslant t\}, \qquad (2.6)$$

где $P\{\xi \geqslant t\}$ — вероятность того, что величина интервала между последним и ожидаемым требованиями больше

или равна t. Легко видеть, что она равна $P_0(t)$. Действительно, $P_0(t)$ — это вероятность того, что за промежуток времени t не придет ни одного требования. Отличие между ними заключается лишь в том, что в первом случае в начальном моменте интервала находится требование, а во втором случае его может не быть. Но так как простейший поток не имеет последействия, расположение требований в предыдущие промежутки времени не имеет значения. В частности, может быть так, что последнее требование из предыдущего промежутка находится в начальной точке рассматриваемого. Следовательно,

$$P\{\xi \geqslant t\} = P_0(t) = e^{-\lambda t}, \qquad (2.7)$$

и из (2.6) и (2.7) получаем (2.1).

— Все ли рекуррентные потоки, как и простейший, можно задавать двумя разными способами? — спросил Борис.

— Посмотрите на формулу (2.5). Из нее видно, что $P_k(t)$ не зависит от того, как чередовались события за время, предшествующее промежутку времени t. В этом также выражается свойство отсутствия последействия. Для описания других потоков Пальма нужно как-то учитывать поведение потока до рассматриваемого промежутка времени. Это удается сделать с помощью введенных А. Я. Хинчиным функций, называемых функциями Пальма — Хинчина, которые аналогичны вероятностям (2.5), но в отличие от них являются условными. Условие заключается в том, что в начальный момент рассматриваемого промежутка времени t пришло требование (оно относится к предыдущему промежутку).

— Александр Андреевич, — сказала Галя, — вы объяснили, почему простейший поток называется стационарным пауссоновским. А что, есть и нестационарный пуассоновский поток?

— Есть. Его задают распределением числа требований, поступивших за промежуток времени t, начало которого — момент t_0. Это тоже распределение Пуассона

$$P_k(t, t_0) = \frac{(\Lambda t)^k}{k!} \cdot e^{-\Lambda t}, \qquad (2.8)$$

но только интенсивность поступления $\lambda(t)$ здесь уже не постоянна, а изменяется во времени. Параметр Λ — это математическое ожидание числа требований, поступивших за отрезок времени t, начало которого расположено в точке t_0.Λ, следовательно, определяется так:

$$\Lambda = \int_{t_0}^{t_0+t} \lambda(\tau)\, d\tau.$$

Поток этот нестационарен и поэтому, конечно, не относится к рекуррентным.

— Давайте вернемся к простейшему потоку, — продолжал Александр Андреевич, — и найдем для него математическое ожидание длины интервала T между моментами поступления требований простейшего потока.

$$M[T] = \int_0^\infty t f(t)\, dt = \lambda \int_0^\infty t e^{-\lambda t}\, dt =$$

$$= -te^{-\lambda t}\Big|_0^\infty + \int_0^\infty e^{-\lambda t}\, dt = \frac{1}{\lambda}. \qquad (2.9)$$

Интенсивность потока

$$\frac{1}{M[T]} = \lambda.$$

Следовательно, параметр λ — это ничто иное, как интенсивность потока, и для того чтобы полностью задать простейший поток, достаточно задать его интенсивность.

Дисперсия длины интервала простейшего потока, как нетрудно убедиться, равна $1/\lambda^2$, а среднее квадратическое отклонение $\frac{1}{\lambda}$.

— Александр Андреевич, — попросила Галя, — расскажите, пожалуйста, о других рекуррентных потоках. Сколько их?

— О, их очень много, — рассмеялся Александр Андреевич, — не пересчитать. Возьмите любую неотрицательную случайную величину и считайте ее функцию распределения функцией распределения интервалов между событиями какого-то рекуррентного потока. Сколько таких распределений — столько и рекуррентных потоков. Конечно, далеко не все из них имеют собственные названия, но о некоторых я могу рассказать.

Мы упоминали уже регулярный поток (с фиксированными интервалами между моментами поступления требований). Он имеет жесткое последействие: если известно, когда пришла последняя заявка, то можно точно сказать, чему равно время ожидания следующей. В этом отношении регулярный поток противоположен простейшему, для которого такая информация не имеет значения.

Иногда встречается нормальный поток — его интервалы распределены по усеченному нормальному закону. Он определен только на положительной полуоси. Так как интервалы между требованиями не могут принимать отрицательных значений, нормальное распределение имеет смысл применять, когда

$$m > 3\sigma, \qquad (2.10)$$

m и σ — математическое ожидание и среднее квадратическое отклонение длины интервала (т. е. вероятность отрицательных значений практически равна нулю).

Если $\sigma \to 0$, то нормальный поток вырождается в регулярный.

Важное место среди рекуррентных потоков занимает семейство потоков Эрланга. Интервалы между моментами появления требований в потоке Эрланга k-го порядка (k может изменяться от 1 до ∞) распределены как суммы k подынтервалов, распределенных по показательному закону (2.1). Это значит, что если мы в простейшем потоке оставим каждое k-е требование, а остальные выбросим, то получим поток Эрланга k-го порядка.

На рис. 2.3 в простейшем потоке оставлено каждое третье требование, остальные выброшены. В результате получен поток Эрланга 3-го порядка.

Плотность распределения интервалов между требованиями в потоке Эрланга k-го порядка (т. е. плотность распределения закона Эрланга k-го порядка) — это композиция k плотностей показательного распределения (2.2).

Она имеет вид:

$$f_k(t) = \begin{cases} \dfrac{\lambda\,(\lambda t)^{k-1}}{(k-1)!}\, e^{-\lambda t} & \text{при } t \geqslant 0 \\ 0 & \text{при } t < 0, \end{cases} \qquad (2.11)$$

где λ — интенсивность простейшего потока.

Если $k=1$, то выражение (2.11) превращается в (2.2).

Обратите внимание на любопытный факт, и мы получаем удивительную закономерность: распределение Эрланга первого порядка — это показательное распределение, а поток Эрланга первого порядка — это простейший поток.

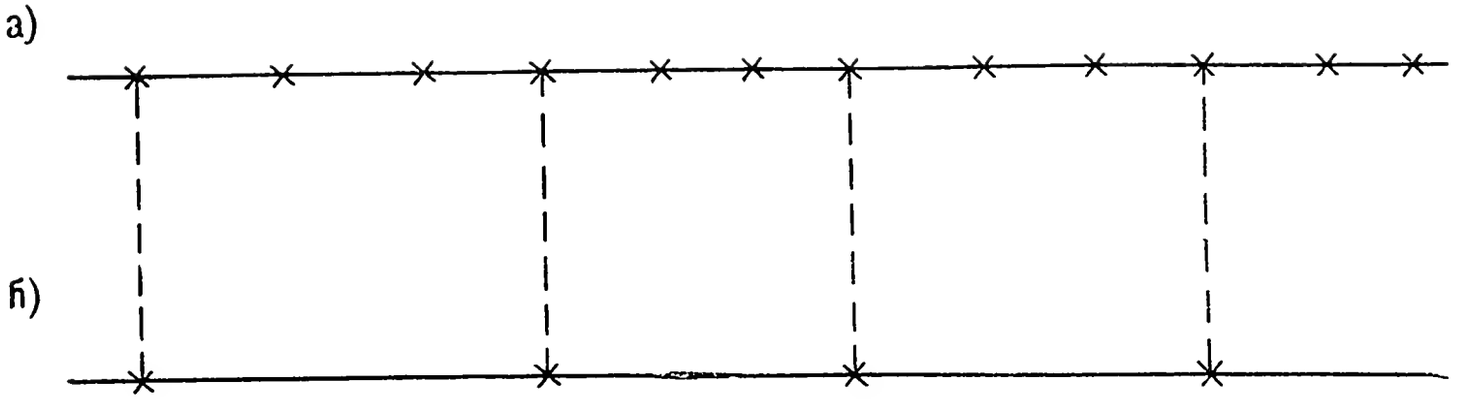


Рис. 2.3

Математическое ожидание и дисперсия закона Эрланга k-го порядка соответственно равны (по теореме сложений математических ожиданий и дисперсий):

$$m_k = k/\lambda \tag{2.12}$$

и

$$D_k = k/\lambda^2. \tag{2.13}$$

Как меняется распределение Эрланга при увеличении k? Из (2.12) видно, что с увеличением порядка (если λ не меняется) интервалы между требованиями увеличиваются, а интенсивность просеянного (эрланговского) потока уменьшается.

Но что будет, если k увеличивается, а интенсивность эрланговского потока (и, следовательно, средняя длина интервалов между требованиями) не меняется? Оказывается, в этом случае при увеличении k распределение интервалов все больше приближается к усеченному нормальному (это хорошо видно на рис. 2.4). При $k \to \infty$ поток становится регулярным. Это можно увидеть из формулы (2.13), если подставить в нее m_k, полученную из (2.12):

$$D_k = m_k^2/k.$$

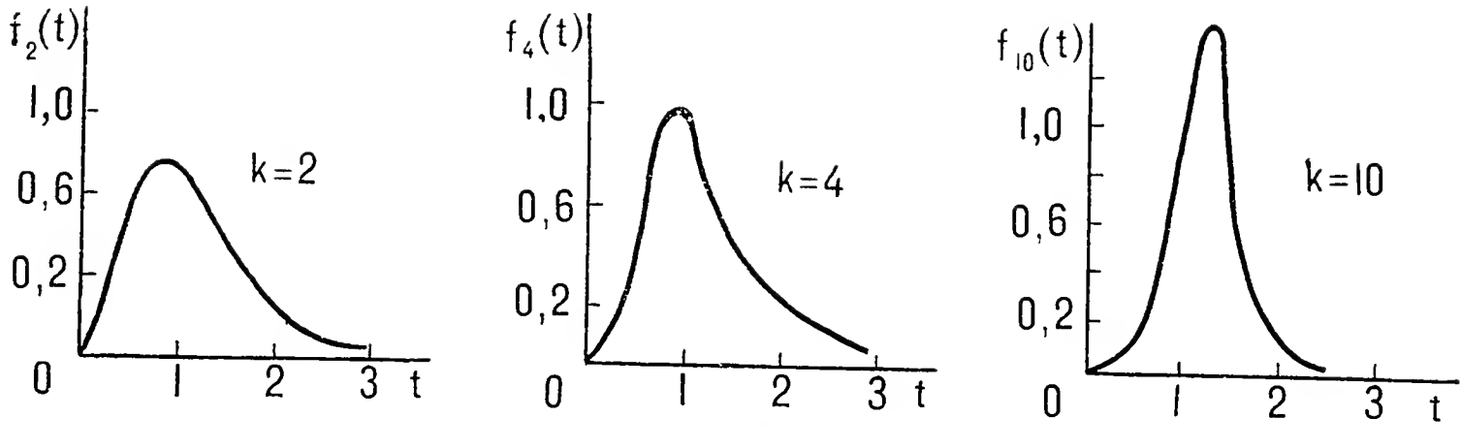


Рис. 2.4

В случае $k \to \infty$, $D_k \to 0$ и интервалы между событиями потока превращаются в одинаковые отрезки времени. Следовательно, степень последействия семейства распределения Эрланга изменяется от отсутствия последействия при $k=1$ до жесткого последействия регулярного потока при $k=\infty$.

Важным параметром потока является коэффициент вариации интервалов между моментами поступления требований. Он определяется как отношение среднего квадратического отклонения σ к математическому ожиданию m длины интервала

$$\nu = \sigma/m.$$

Для простейшего потока среднее квадратическое отклонение равно математическому ожиданию длины интервалов между поступлениями требований, поэтому коэффициент вариации $\nu=1$. У всех других потоков, о которых мы говорили, коэффициент вариации меньше единицы. Например, у потока Эрланга k-го порядка $\nu=1/\sqrt{k}$. В частности, эрланговский поток второго порядка имеет коэффициент вариации $\nu=0{,}71$. Чем больше порядок k, тем меньше коэффициент вариации. В предельном случае, при $k \longrightarrow \infty$ получается регулярный, детерминированный поток, у которого, как видно, коэф-

фициент вариации равен нулю. Таким образом, коэффициент вариации — это параметр, характеризующий степень неравномерности поступления требований.

— А бывают потоки, у которых коэффициент вариации больше единицы?

— Бывают потоки со сколь угодно большими коэффициентами вариации. Но на практике мы почти всегда имеем дело с потоками, у которых коэффициенты вариации не превосходят единицы. Дело в том, что степень неравномерности потоков, которые мы обычно наблюдаем, как правило, не больше степени неравномерности простейшего потока. Потоки с коэффициентами вариации больше единицы имеют очень высокую степень неравномерности, которая увеличивается с ростом коэффициента вариации. Но поскольку это вас интересует, я приведу в качестве примера поток с гиперэкспоненциальным распределением второго порядка интервалов между моментами поступления требований. Это распределение имеет плотность

$$f(t)=\begin{cases} p\lambda_1 e^{-\lambda_1 t}+(1-p)\lambda_2 e^{-\lambda_2 t} & \text{при } t\geqslant 0 \\ 0 & \text{при } t\leqslant 0, \end{cases}$$

где $0<p<1$. Такое распределение появляется, когда с вероятностью p берут случайную величину, распределенную по показательному закону с параметром λ_1, а с вероятностью $(1-p)$ — случайную величину, распределенную по показательному закону с параметром λ_2. Коэффициент вариации потока с таким распределением интервалов больше единицы.

Теперь, познакомившись с тем, как требования прибывают в систему, посмотрим, как происходит обслуживание требований.

Одна из важнейших характеристик, связанных с работой системы, — длительность обслуживания одного

требования (назовем ее интервалом обслуживания) — в общем случае — случайная величина. Мы будем предполагать (так же, как для входящих потоков), что интервалы обслуживания — взаимно независимые величины, распределенные по одному закону. Этот закон может быть показательным, эрланговским, усеченным нормальным (свойства этих распределений мы уже рассмотрели и все сказанное о них применительно к интервалам входящего потока целиком относится и к интервалам обслуживания) или каким-нибудь иным. Интервал обслуживания может быть и неслучаен, если время обслуживания одно и то же для всех требований. Постоянный интервал обслуживания имеет такая система обслуживания, как, например, фотоавтомат.

Интересно, что показательное распределение интервалов обслуживания встречается сравнительно реже, чем показательное распределение интервалов входящего потока.

— Чем это объясняется? — поинтересовалась Галя.

— Тем, что очень часто существует некоторый промежуток времени, раньше которого обслуживание не может закончиться. Например, корабль в порту или вагон на грузовой станции не могут быть разгружены мгновенно, поезд не может быть мгновенно расформирован.

Из рис. 2.2 видно, что при показательном распределении короткие интервалы имеют значительный вес, поэтому наложенное только что ограничение не позволяет распределению интервалов обслуживания быть показательным. А вот телефонный разговор может быть сколь угодно коротким и распределение длительности разговоров хорошо описывается показательным законом. Иногда, хотя и реже, минимальный интервал бывает и у входящих потоков. Например, поток поездов, идущих в одном направлении, имеет такой интервал, и распре-

деление промежутков времени между поездами оказывается непоказательным.

— Значит, поток поездов, поступающих на нашу сортировочную станцию, имеет минимальный интервал между моментами поступлений. Что же это за поток? — спросила Галя.

— Вы не учли, что на сортировочную станцию поступают поезда с нескольких направлений, — сказал Стрелкин, — хотя по каждому направлению имеется минимальный интервал, с разных направлений поезда могут прийти почти одновременно.

— Здесь как раз происходит суммирование или наложение потоков, — сказал Александр Андреевич.

— И наверно, в результате получается простейший поток? — спросил Стрелкин.

— Не получается, — сказал Боря. — На лекциях нам говорили, что проводились исследования таких потоков. Тогда я не знал теории вероятностей и мало что понял. Теперь я, представляю себе, что происходит. Хотя здесь действительно суммируется несколько потоков, но, как правило, два встречных потока главного направления имеют значительно бóльшую интенсивность, чем остальные. А как говорил Александр Андреевич, для того чтобы небольшое число суммируемых потоков давало хорошее приближение к простейшему, надо, чтобы их интенсивности были соизмеримыми.

— Какой же поток все-таки получается?

— Оказалось, что хорошую аппроксимацию дают иногда эрланговские потоки невысоких порядков ($k=2$). В целом же коэффициент вариации интервалов прибытия поездов колеблется в пределах 0,7—0,9.

— Наверно, исследовались и характеристики обслуживания? — спросил Александр Андреевич.

— Исследовались. Оказалось, что длительность технического осмотра удобно считать распределенной по

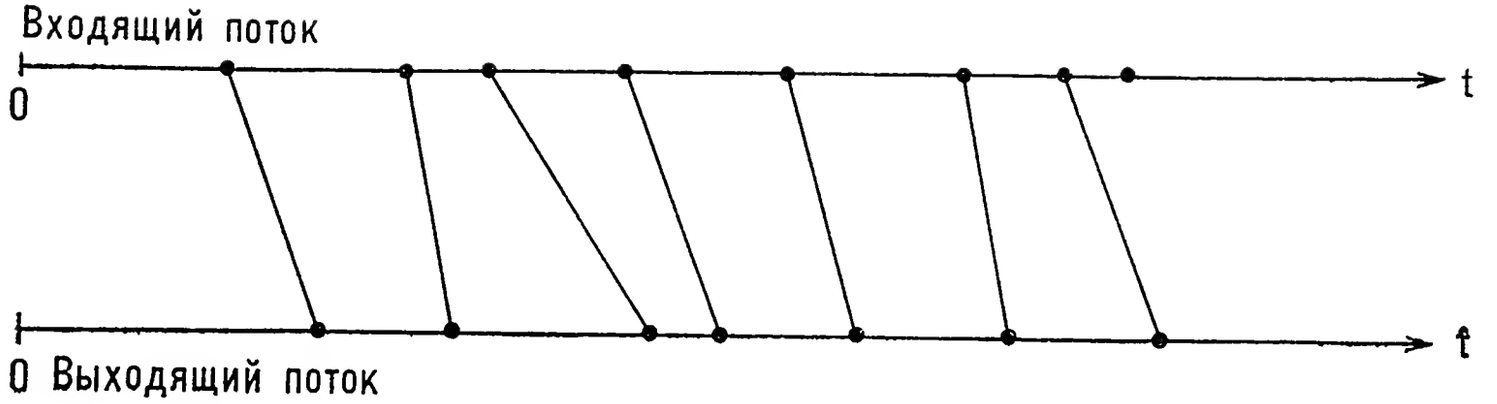


Рис. 2.5

усеченному нормальному закону или по закону Эрланга высокого порядка ($k=15$—16).

— Александр Андреевич, — спросила Галя, — а что получается после обслуживания, что представляет собой выходящий поток и насколько он отличается от входящего?

— Процесс обслуживания — это некоторое преобразование входящего потока (рис. 2.5). Очень часто важно знать характеристики выходящего потока, полученного в результате такого преобразования. Дело в том, что значительное место среди систем массового обслуживания занимают многофазовые системы, в которых требования после обслуживания на первой фазе поступают на вторую, потом, может быть, на третью и так далее. В этом случае выходящий поток первой фазы — это входящий поток второй. В рассмотренном нами примере выходящий после технического осмотра поток составов является входящим потоком для второй фазы обслуживания — расформирования на горке.

Что же представляет собой выходящий поток? Сразу заметим, что его интервалы — не то же самое, что интервалы обслуживания. Действительно, может быть так, что после окончания очередного обслуживания система какое-то время не работает из-за отсутствия требований. В этом случае интервал выходящего потока состо-

ит из периода незанятости системы и интервала обслуживания первого пришедшего после простоя требования, т. е., если $t_n^{\text{обсл}}$ — интервал обслуживания n требования, а $t_n^{\text{вых}}$ — интервал между моментами выхода n-го и $n+1$-го требований, то

$$t_n^{\text{обсл}} \leqslant t_n^{\text{вых}} . \qquad (2.14)$$

Распределение интервалов потока, выходящего из одноканальной системы, легко получить в двух крайних случаях: когда система работает с большой нагрузкой и когда система работает с малой нагрузкой. Если нагрузка велика, периодов простоя практически нет и интервалы выходящего потока почти всегда совпадают с интервалами обслуживания. Если нагрузка мала, требования приходят в систему редко, а обслуживаются быстро, моменты выхода требований из системы — это моменты входа, сдвинутые на случайную величину, — продолжительность обслуживания.

Нас, естественно, интересует промежуточный случай, когда интервалы входящего потока и обслуживания — величины одного порядка и в то же время имеются периоды простоя системы.

Оказывается, что если в систему (одноканальную или многоканальную) поступает простейший поток требований и время обслуживания имеет показательное распределение, выходящий поток также является простейшим.

Что же будет, если время обслуживания имеет непоказательное распределение? В этом случае для одноканальной системы оказывается, что выходящий поток не только не простейший, но не является даже потоком с ограниченным последействием, т. е. его интервалы — взаимно зависимые величины. Часто, правда, эта зависимость оказывается довольно слабой, и тогда выхо-

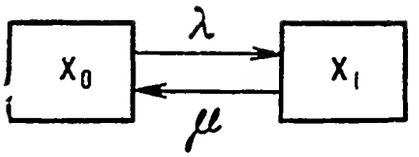


Рис. 2.6

дящий поток можно с известной долей приближения считать рекуррентным.

Мы уже упоминали системы массового обслуживания с отказами, где пришедшее требование, застав все приборы занятыми, уходит необслуженным. Поток необслуженных заявок может поступать на обслуживание в другую систему, и в таких случаях нам интересно знать, каковы свойства этого потока?

Так вот, если в многоканальную систему с отказами и показательно распределенным временем обслуживания поступает рекуррентный поток, то поток необслуженных требований в системе является рекуррентным.

Интересно, что в отличие от предыдущего случая, если входящий поток простейший, поток необслуженных требований простейшим не будет.

— Итак, — сказал Александр Андреевич, — мы знаем, как требования поступают в системы и как они обслуживаются. Теперь мы рассмотрим некоторые наиболее интересные системы массового обслуживания и найдем их характеристики.

В случайные моменты времени происходят события, которые переводят систему из одного состояния в другое: приходит новое требование, обслуженное требование покидает систему, выходит из строя обслуживающее устройство и т. д. Потоки всех возможных событий определяют случайный процесс, протекающий в системе массового обслуживания. На рис. 2.6 показаны два состояния, в которых может находиться одноканальная система без очереди. x_0 означает, что заявок в системе нет и прибор свободен, x_1 — пришло требование и прибор занят его обслуживанием. Если в это время придет новое требование, оно будет потеряно.

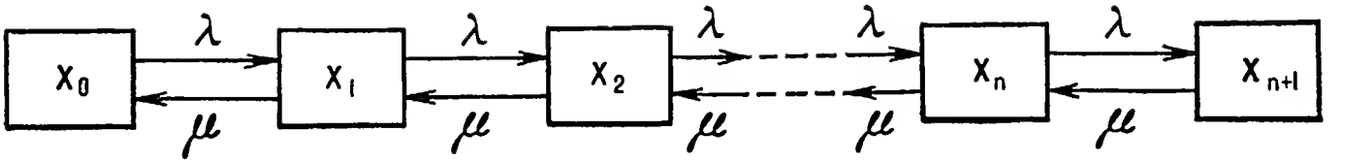


Рис. 2.7

Когда обслуживание будет закончено, система вернется в состояние x_0 и дальше процесс будет продолжаться таким же образом. λ и μ — это соответственно интенсивность входящего потока требований и скорость обслуживания.

— Что такое скорость обслуживания? — спросил Борис.

— Выражение «скорость обслуживания» действительно нужно пояснить. Когда определяют среднюю скорость автомобиля, в расчет принимают время движения и время стоянок (если они были). Но можно считать и среднюю скорость движения, при определении которой учитывают только «чистое» время движения автомобиля. Точно так же при определении скорости обслуживания учитывается только время работы обслуживающего устройства и не учитывается время простоев.

Мы будем считать, что входящий поток является рекуррентным потоком, а продолжительности обслуживания — независимые (между собой и от интервалов входящего потока) и одинаково распределенные величины. Требования приходят и обслуживаются по одному.

Рассмотрим несколько обобщений этой простейшей системы.

Одноканальная система с ограниченной очередью (рис. 2.7). Система имеет n мест для ожидания. Если имеется хотя бы одно свободное место, пришедшее требование поступает в систему; если свободных мест нет, требование получает отказ и теряется. В системе одновременно могут находиться не более чем $n+1$ требова-

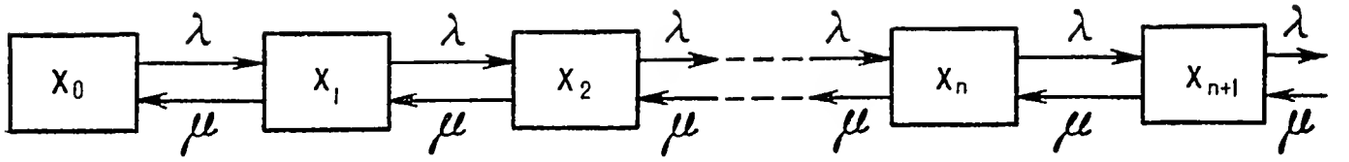


Рис. 2.8.

ний: n требований в очереди и одно на обслуживании. Система имеет $n+2$ состояний.

Одноканальная система с неограниченной очередью (рис. 2.8). Число мест в очереди не ограничено, всякое пришедшее требование попадает в систему. Число состояний системы бесконечно, однако их можно «пересчитать»: x_0, x_1, x_2, x_3. Такое бесконечное множество называется счетным.

Многоканальная система без очереди (рис. 2.9). Система имеет n одинаковых обслуживающих приборов. Если хотя бы один обслуживающий прибор свободен, пришедшее требование немедленно поступает на обслуживание. Если все приборы заняты, пришедшее требование теряется. Такая система имеет $n+1$ состояний. Заметим, что если одновременно работают два прибора, скорость обслуживания увеличивается в 2 раза и равна 2μ, а если одновременно работают n приборов, скорость обслуживания $n\mu$.

Система с бесконечным числом приборов (рис. 2.10). Для любого требования, пришедшего в эту систему, всегда найдется свободный прибор. Количество работающих приборов не ограничено. Число состояний системы — счетно.

Многоканальная система с ограниченной очередью (рис. 2.11). Система имеет n обслуживающих приборов

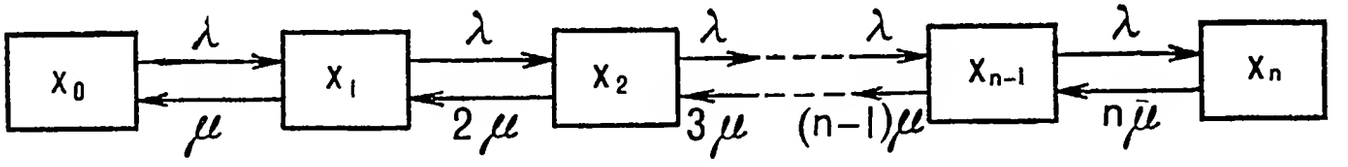


Рис. 2.9.

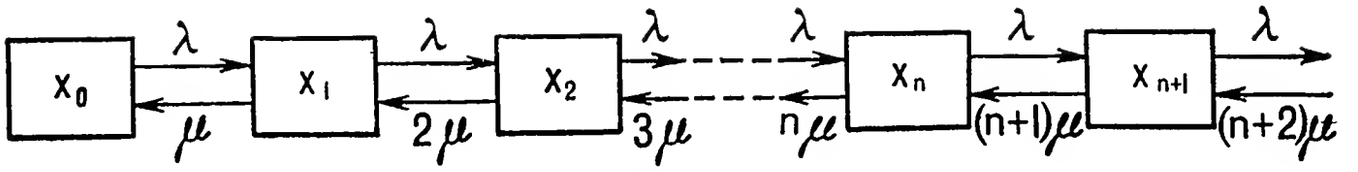


Рис. 2.10

и m мест для ожидания. Одновременно в системе может находиться не более чем $n+m$ требований. Система может находиться в $n+m+1$ состояниях.

Многоканальная система с неограниченной очередью (рис. 2.12). Эта система отличается от предыдущей тем, что число мест в очереди не ограничено и число состояний счетно.

Эти системы можно усложнить (например, учитывать возможность выхода приборов из строя), однако мы ограничимся рассмотрением лишь перечисленных.

При исследовании системы массового обслуживания важно найти вероятности, с которыми система находится в каждом из состояний. В общем случае эти вероятности зависят от времени, прошедшего с момента начала работы системы. Дело в том, что после начала работы в системе происходит переходный процесс. Мы будем рассматривать системы массового обслуживания только в стационарном режиме, т. е. будем считать, что система работает достаточно долго и вероятности состояний системы уже не зависят от времени (т. е. от момента начала работы). Эти величины называются предельными вероятностями.

Следует сказать, что не любая система имеет предельные вероятности состояний. Для системы с неограниченной очередью их существование зависит от того,

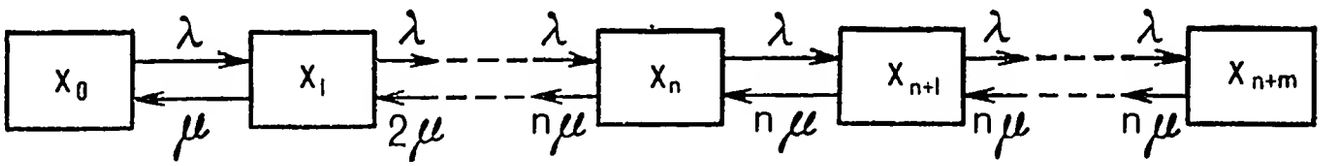


Рис. 2.11

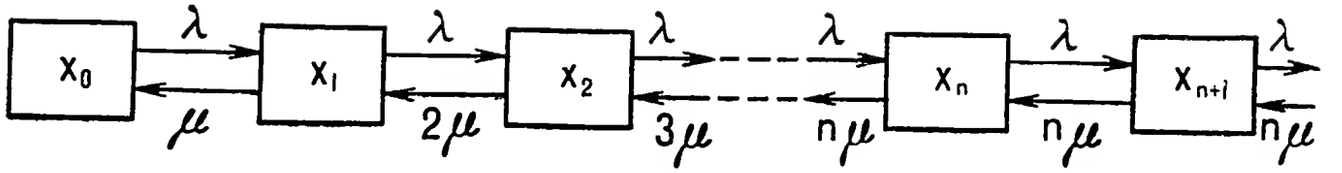


Рис. 2.12

чему равна нагрузка системы. Нагрузкой называется отношение величины интенсивности входящего потока к произведению скорости обслуживания одного прибора на число приборов. Для одноканальных систем нагрузка $\rho = \frac{\lambda}{\mu}$. Для n-канальных систем $\rho = \frac{\lambda}{n\mu}$.

Все системы, которые мы рассмотрим, имеют предельные вероятности состояний. При этом для систем с неограниченной очередью требуется выполнение условия $\rho < 1$.

— У этого условия есть какой-нибудь физический смысл? — спросил Борис.

— Конечно, оно означает, что скорость обслуживания больше скорости поступления требований. В случае его невыполнения очередь будет расти до бесконечности. В дальнейшем мы будем считать, что для систем с неограниченной очередью $\rho < 1$.

Если входящий поток требований — простейший, а продолжительность обслуживания одного требования распределена по показательному закону, то проходящий в системе массового обслуживания процесс будет марковским.

Процесс называется марковским, если его поведение в будущем зависит только от состояния системы в настоящий момент и не зависит от того, как она попала в это состояние. То, что такая система будет марковской, легко следует из доказанного нами свойства отсутствия последействия простейшего потока. Действительно, мы показали, что если интервалы между событиями распределены по показательному закону, длина

оставшегося интервала в какой-то момент времени до наступления очередного события распределена по тому же показательному закону и не зависит от того, что было до этого момента. Таким образом, если интервалы между поступлениями требований в систему и продолжительности обслуживания распределены по показательным законам, то вероятность любого состояния системы в будущем зависит только от состояния системы в настоящий момент.

Для марковских систем массового обслуживания легко можно написать систему алгебраических уравнений, линейных относительно предельных вероятностей состояний с помощью следующего мнемонического правила: *предельная вероятность i-го состояния, умноженная на сумму интенсивностей потоков, переводящих систему в другие состояния, равна сумме произведений предельных вероятностей состояний, из которых можно перейти в i-е за один шаг, на интенсивности этих переходов.*

Такое уравнение можно написать для каждого состояния. К полученной системе надо добавить уравнение — нормировочное условие

$$\sum_{i=0}^{\infty} p_i = 1. \tag{2.15}$$

Для того чтобы написать такую систему уравнений, достаточно иметь граф состояний. Например, марковская система, граф состояний которой изображен на рис. 2.6, описывается системой уравнений:

$$\begin{cases} \lambda p_0 = \mu p_1; \\ \mu p_1 = \lambda p_0; \\ p_0 + p_1 = 1. \end{cases}$$

— Но ведь первые два уравнения одинаковы, — заметила Галя.

— Да, из них достаточно оставить одно:

$$\begin{cases} \lambda p_0 = \mu p_1; \\ p_0 + p_1 = 1. \end{cases} \tag{2.16}$$

Откуда

$$p_0 = \frac{\mu}{\lambda + \mu}; \tag{2.17}$$

$$p_1 = \frac{\lambda}{\lambda + \mu}, \tag{2.18}$$

где p_0 — это вероятность того, что система обслуживания свободна;

p_1 — вероятность того, что прибор занят, т. е. вероятность того, что пришедшее требование будет потеряно.

Рассмотрим теперь более сложный случай — многоканальную систему без очереди. Имеется n одинаковых каналов, продолжительность обслуживания одного требования распределена по показательному закону с параметром μ, в систему поступает простейший поток требований с интенсивностью λ. Требование, поступившее, когда все приборы заняты, покидает систему. Каждый прибор может одновременно обслуживать только одно требование.

Основными параметрами, характеризующими работу такой системы, являются вероятность отказа, т. е. вероятность занятости всех приборов в момент поступления очередного требования, и среднее число занятых приборов.

Вероятность отказа характеризует отношение количества потерянных требований за какое-то время к числу всех поступивших за это время требований. Среднее

число занятых приборов — это характеристика степени загрузки обслуживающей системы.

С помощью графа состояний (см. рис. 2.9) и мнемонического правила запишем систему уравнений для предельных вероятностей:

$$\begin{cases} \lambda p_0 = \mu p_1; \\ (\lambda + \mu) p_1 = \lambda p_0 + 2\mu p_2; \\ \cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots \\ (\lambda + \mu) p_k = \lambda p_{k+1} + (k+1)\mu p_{k+1} \quad (0 < k < n); \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ n\mu p_n = \lambda p_{n-1}; \\ \sum\limits_{k=0}^{n} p_k = 1 \end{cases} \tag{2.19}$$

Из первого уравнения получаем

$$p_1 = \frac{\lambda}{\mu} p_0. \tag{2.20}$$

Из второго

$$p_2 = \frac{1}{2\mu}[(\lambda + \mu) p_1 - \lambda p_0] = \frac{\lambda^2}{2\mu^2} p_0. \tag{2.21}$$

Продолжая дальше, получим:

$$p_3 = \frac{\lambda^3}{3!\mu^3} p_0,$$

$$p_4 = \frac{\lambda^4}{4!\mu^4} p_0$$

И вообще, для $k \leqslant n$

$$p_k = \frac{\lambda^k}{k!\mu^k} p_0. \tag{2.22}$$

В соответствии с последним уравнением системы (2.19):

$$\sum_{k=0}^{n} p_k = p_0 \sum_{k=0}^{n} \frac{\lambda^k}{k!\mu^k} = 1, \tag{2.23}$$

откуда вероятность того, что в системе нет ни одного требования, равна

$$p_0 = \frac{1}{\sum_{k=0}^{n} \frac{\lambda^k}{k!\mu^k}}. \tag{2.24}$$

Если значение p_0 подставить в (2.22), получим так называемые формулы Эрланга

$$p_k = \frac{\frac{\lambda^k}{k!\mu^k}}{\sum_{s=0}^{n} \frac{\lambda^s}{s!\mu^s}} \quad (0 \leqslant k \leqslant n). \tag{2.25}$$

Несмотря на то, что формулы Эрланга были выведены в предположении, что время обслуживания распределено по показательному закону, они, как было впоследствии доказано, справедливы для любого закона обслуживания. Таким образом, вероятности состояний системы зависят только от среднего времени обслуживания требования одним каналом $(1/\mu)$ и не зависят от вида распределения продолжительности обслуживания.

В случае занятости всех каналов требование не принимается на обслуживание и получает отказ. Вероятность отказа по формуле (2.25):

$$p_n = \frac{\frac{\lambda^n}{n!\mu^n}}{\sum_{s=0}^{a} \frac{\lambda^s}{s!\mu^s}}, \qquad (2.26)$$

а среднее число занятых приборов равно

$$M[k] = \sum_{k=1}^{n} k p_k. \qquad (2.27)$$

— Может быть, посчитаем какой-нибудь пример? — предложил Стрелкин.

— Давайте, — согласился Александр Андреевич. Назовем его:

Пример 1. Около поста ГАИ три автоинспектора проверяют путевые листы у водителей грузовых автомобилей. Если хотя бы один инспектор свободен, проезжающий грузовик останавливают, если все инспектора заняты, грузовик проезжает мимо. Поток грузовиков простейший, его интенсивность равна одному грузовику в минуту.

На проверку путевого листа у инспектора уходит в среднем 5 мин. У какой части водителей путевой лист не будет проверен?

Легко видеть, что в этом примере описана трехканальная система массового обслуживания без очереди.

Интенсивность входящего потока $\lambda = 1$, скорость обслуживания одного «прибора» $\mu = \frac{1}{5}$. Интересующая

нас величина может быть вычислена по формуле Эрланга для $n=3$:

$$p_3 = \frac{\frac{\lambda^3}{3!\mu^3}}{\sum_{s=0}^{3}\frac{\lambda^s}{s!\mu^s}} = \frac{\frac{1}{3!\left(\frac{1}{5}\right)^3}}{\sum_{s=0}^{3}\frac{1}{s!\left(\frac{1}{5}\right)^5}} \approx 0{,}53.$$

Таким образом, более чем у половины водителей путевой лист проверен не будет.

Если число приборов бесконечно, то система уже не может рассматриваться как система с отказами, так как каждое требование будет обслужено. На практике таких систем, конечно, нет, но в случаях, когда приборов так много, что потерь практически не бывает, удобно считать число приборов бесконечным. Такие системы встречаются в ситуациях, когда требования ожидать не могут, а потери недопустимы.

Систему с бесконечным числом приборов можно рассматривать, как предельный случай n-канальной системы (граф состояний показан на рис. 2.10). Если в формулах Эрланга перейти к пределу при $n\to\infty$, получим, что вероятности состояний в стационарном режиме распределены по закону Пуассона:

$$p_k = e^{-\frac{\lambda}{\mu}}\frac{\lambda^k}{k!\mu^k} \quad (k=0,\ 1,\ 2,\ldots). \qquad (2.28)$$

Так же, как и для случая конечного числа приборов, эти формулы справедливы при любом распределении длительности обслуживания.

Мы уже говорили о том, что в многоканальной системе с неограниченной очередью, в которой входящий

поток — простейший, а обслуживание — показательное, выходящий поток тоже простейший. В системе с бесконечным числом приборов при тех же условиях выходящий поток также оказывается простейшим. Более того, если время обслуживания распределено по произвольному закону, выходящий поток, тем не менее, — простейший.

— Теперь поговорим о системах с очередью, — предложил Александр Андреевич.

— Требование, поступившее в систему с неограниченной очередью, не уйдет, пока не будет обслужено. Как и прежде, мы будем рассматривать системы в стационарном режиме. Основные параметры таких систем — это величины, характеризующие количество требований в системе, и время нахождения требований в системе. К первым относятся среднее число требований в системе ($L_{\text{сист}}$ — здесь считают все требования — и те, которые обслуживаются, и те, которые стоят в очереди) и средняя длина очереди ($L_{\text{оч}}$ — здесь учитываются только требования, ожидающие обслуживания). Ко вторым — среднее время нахождения требования в системе ($W_{\text{сист}}$) и среднее время ожидания требования в очереди ($W_{\text{оч}}$). Нетрудно заметить, что

$$W_{\text{сист}} - W_{\text{оч}} = \frac{1}{\mu}, \tag{2.29}$$

где $1/\mu$ — средняя продолжительность обслуживания одного требования.

Существует очень важная формула, связывающая среднее число требований со средним временем ожидания в произвольной системе массового обслуживания с неограниченной очередью

$$L = \lambda W, \tag{2.30}$$

справедливая как для $L_{\text{сист}}$ и $W_{\text{сист}}$, так и для $L_{\text{оч}}$ и $W_{\text{оч}}$ (λ — это интенсивность входящего потока). Эта простая на вид формула была доказана лишь в 1961 г. американским математиком Литлом и носит его имя. Доказательство Литла было громоздким и впоследствии оно было упрощено. Мы покажем очень простое доказательство этой формулы.

Верхняя линия рис. 2.13 описывает общее число требований, пришедших в систему, а нижняя — число требований, ушедших из нее. Вертикальный разрез соответствует числу требований в системе.

Рассмотрим период времени T в стационарном режиме (T может включать несколько периодов занятости) и введем следующие обозначения:

$Z(T)$ — общее число прибытий за период T;

$A_{\text{сист}}(T)$ — общее время нахождения требований в системе за период T, равное заштрихованной площади на рис. 2.13;

$\lambda(T)$ — средняя интенсивность прибытий за T, равная $Z(T)/T$;

$W_{\text{сист}}(T)$ — среднее время нахождения в системе за T, равное $A_{\text{сист}}(T)/Z(T)$;

$L_{\text{сист}}(T)$ — среднее число требований в системе за T, равное $A_{\text{сист}}(T)/T$.

Сопоставляя выражения для параметров системы: $\lambda(T)$, $W_{\text{сист}}(T)$ и $L_{\text{сист}}(T)$, видно, что

$$L_{\text{сист}}(T) = \lambda(T) W_{\text{сист}}(T).$$

Это доказательство не зависит от каких-либо предположений относительно распределений интервалов прибытия и обслуживания, о числе приборов и дисциплине обслуживания.

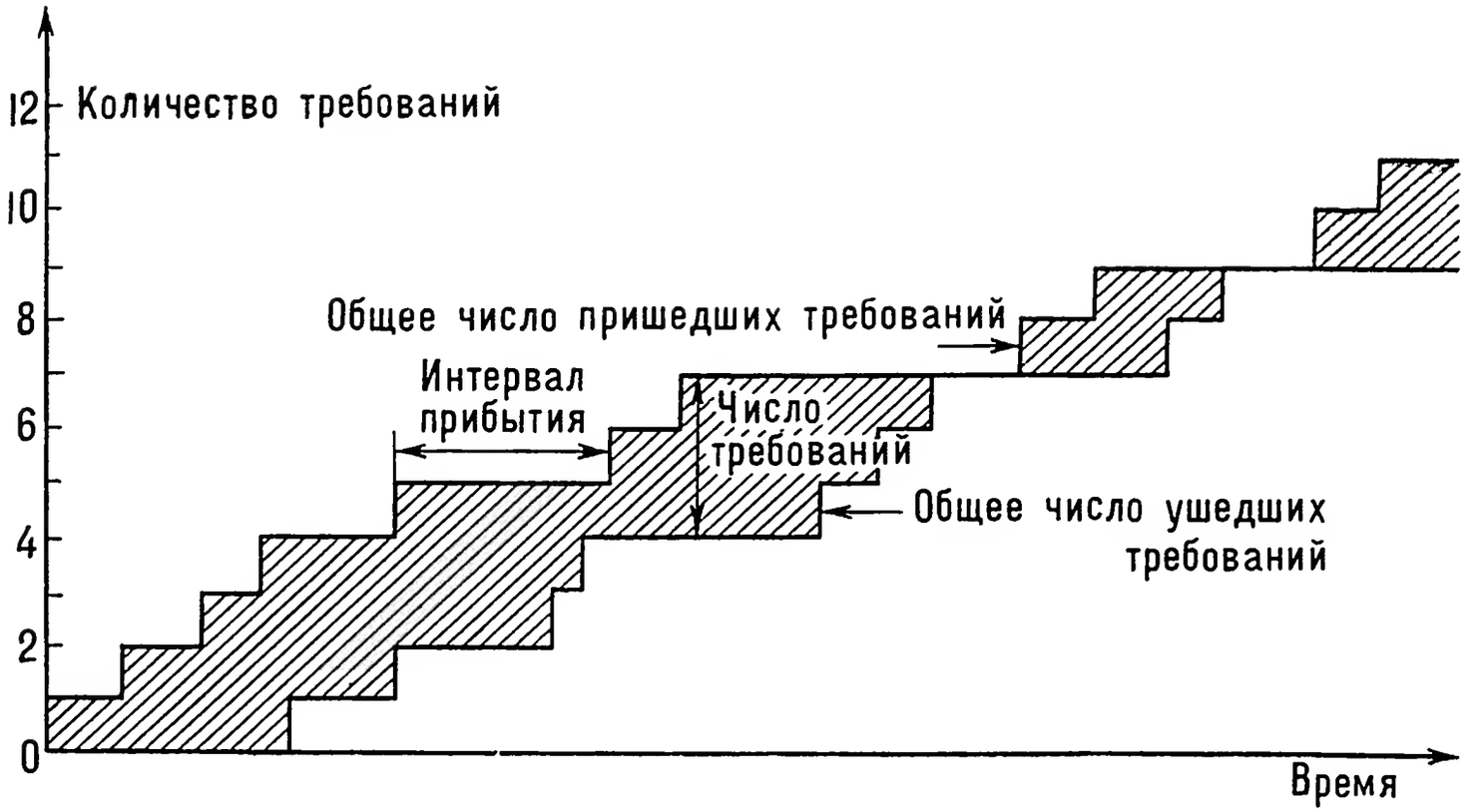


Рис. 2.13

Дадим определение пределов при $T \to \infty$. Если существуют

$$\lambda = \lim_{T \to \infty} \lambda(T) \quad \text{и} \quad W_{\text{сист}} = \lim_{T \to \infty} W_{\text{сист}}(T),$$

то предел $L_{\text{сист}}(T)$ также существует, — определим его

$$L_{\text{сист}}(T) = \lim_{T \to \infty} L_{\text{сист}}(T),$$

и все пределы должны удовлетворять отношению

$$L_{\text{сист}} = \lambda W_{\text{сист}}. \tag{2.31}$$

Точно так же, если нижняя линия на рис. 2.13 соответствует уходам из очереди, а не из системы, получается аналогичный результат

$$L_{\text{оч}} = \lambda W_{\text{оч}}. \tag{2.32}$$

Рассмотрим систему массового обслуживания для случая одного прибора (см. рис. 2.8), простейшего входящего потока с интенсивностью λ и длительностью обслуживания, распределенной по показательному закону с параметром μ.

С помощью мнемонического правила составим систему уравнений для стационарного режима:

$$\begin{cases} \lambda p_0 = \mu p_1; \\ (\lambda + \mu) p_1 = \lambda p_0 + \mu p_2; \\ (\lambda + \mu) p_n = \lambda p_{n-1} + \mu p_{n+1} \end{cases} \tag{2.33}$$

и

$$\sum_{n=0}^{\infty} p_n = 1. \tag{2.34}$$

Последовательно решая уравнения системы (2.33), найдем

$$p_n = \rho^n p_0 . \tag{2.35}$$

Напомним, что для одноканальных систем, как нами замечено ранее, $\rho = \lambda/\mu$.

Подставляя эти выражения в (2.34), получим

$$p_0 \sum_{n=0}^{\infty} \rho^n = 1, \tag{2.36}$$

откуда

$$p_0 = 1 - \rho \tag{2.37}$$

и

$$p_n = (1 - \rho)\rho^n . \tag{2.38}$$

Заметим, что выражение (2.37) справедливо не только для рассматриваемой системы, но вообще для

всех одноканальных систем с неограниченной очередью, у которых $\rho<1$.

Найдем среднее число требований в системе

$$L_{\text{сист}}=\sum_{n=0}^{\infty} np_n=\rho(1-\rho)\sum_{n=1}^{\infty} n\rho^{n-1}. \qquad (2.39)$$

Нетрудно показать, что

$$\sum_{n=1}^{\infty} n\rho^{n-1}=1/(1-\rho)^2. \qquad (2.40)$$

Вместо $\sum_{n=1}^{\infty} n\rho^{n-1}$ рассмотрим величину $\sum_{n=1}^{\infty} nx^{n-1}$ при $0<x<1$.

Очевидно, она является производной от $\sum_{n=1}^{\infty} x^n$, т. е.

$$\sum_{n=1}^{\infty} nx^{n-1}=\left(\sum_{n=1}^{\infty} x^n\right)'=\left(\frac{x}{1-x}\right)'. \qquad (2.41)$$

Заменяя x на ρ, получим выражение (2.40). Подставив (2.40) в (2.39), получим

$$L_{\text{сист}}=\frac{\rho}{1-\rho}. \qquad (2.42)$$

Из (2.42) легко получить остальные средние характеристики системы. $L_{\text{сист}}$ и $L_{\text{оч}}$ отличаются на ρ, так как математическое ожидание числа находящихся на приборе требований равно

$$0(1-\rho)+1\rho=\rho.$$

Отсюда

$$L_{\text{оч}}=L_{\text{сист}}-\rho=\frac{\rho^2}{1-\rho}. \qquad (2.43)$$

Средние временны́е характеристики найдем по формулам (2.31) и (2.32).

$$W_{\text{сист}} = L_{\text{сист}} / \lambda = 1/\mu(1-\rho); \tag{2.44}$$

$$W_{\text{оч}} = L_{\text{оч}}/\lambda = \rho/\mu(1-\rho). \tag{2.45}$$

Мы видим, что выражения для средних характеристик одноканальной системы с неограниченной очередью в случае простейшего входящего потока и показательного распределения длительности обслуживания очень просты. К сожалению, далеко не всегда в конкретных практических задачах удается аппроксимировать распределения длин интервалов входящего потока и обслуживания показательными законами. Тем не менее, даже в таких случаях формулы (2.42)—(2.45) могут оказаться полезными. Эти формулы можно обобщить и на случай n-канальной системы с неограниченной очередью, простейшим входящим потоком и показательно распределенной длительностью обслуживания (см. рис. 2.12). Мы не будем приводить вывод выражений для вероятностей состояний и средних характеристик этой системы, а дадим окончательные результаты.

Вероятность того, что в системе находятся k требований:

$$p_k = \begin{cases} \dfrac{1}{k!}\left(\dfrac{\lambda}{\mu}\right)^k p_0, & 1 < k \leqslant n; \\ \dfrac{1}{n!\,n^{k-n}}\left(\dfrac{\lambda}{\mu}\right)^k p_0, & k > n; \end{cases} \tag{2.46}$$

$$p_0 = \left[\sum_{k=0}^{n-1} \frac{(k\rho)^k}{k!} + \frac{(n\rho)^n}{n!(1-\rho}\right]^{-1}. \tag{2.47}$$

Напомним, что для n-канальной системы $\rho = \lambda/n\mu$.

Среднее число требований в системе и в очереди равно соответственно

$$L_{\text{сист}} = \frac{\rho (n\rho)^n}{n!\,(1-\rho)^2}\, p_0 + n\rho; \qquad (2.48)$$

$$L_{\text{оч}} = \frac{\rho (n\rho)^n}{n!\,(1-\rho)^2}\, p_0. \qquad (2.49)$$

Среднее время нахождения требования в системе и ожидания легко получить с помощью формулы Литла.

— Давайте проиллюстрируем это примером 2 (продолжение примера 1). Заметив, что бóльшая часть грузовиков проезжает без проверки, инспектора решили изменить тактику и проверять все грузовики. Но они быстро подсчитали, что так как в среднем каждую минуту подъезжает новый грузовик, а проверить они смогут в среднем не больше $3 \times 1/5$ путевых листов, то очень скоро около пункта ГАИ вырастет длинный хвост машин. Поэтому они решили работать вдвое быстрее и просматривать путевой лист за 2,5 мин. Сколько в среднем грузовиков стояло около пункта ГАИ и сколько времени в среднем стоял один грузовик?

В этом случае $\rho = \frac{2,5}{3} = \frac{5}{6}$. По формуле (2.48):

$$L_{\text{сист}} = \frac{3^3 \cdot \left(\frac{5}{6}\right)^4}{3!\left(1-\frac{5}{6}\right)^2} \cdot \frac{1}{20} + 3 \cdot \frac{5}{6} \approx 4,45.$$

Среднее время стоянки одного грузовика

$$W_{\text{сист}} = \frac{L_{\text{сист}}}{\lambda} = \frac{4,45}{1} = 4,45 \text{ мин.}$$

— Рассмотрим теперь очереди с ограниченным числом мест для ожидания, — продолжил Александр Андреевич. — Система состоит из n обслуживающих приборов и очереди, длина которой не может превышать m. Если вновь пришедшее требование застает в системе $n+m$ требований (n на обслуживании и m в очереди), то оно получает отказ и теряется. Схема состояний системы показана на рис. 2.11. По-прежнему будем считать, что в систему поступает простейший поток требований с интенсивностью λ, продолжительность обслуживания одного требования распределена по показательному закону с параметром μ.

Не будем выписывать систему уравнений для стационарного режима и приводить ее решение — все это аналогично тому, что мы проделали для системы без очереди. Приведем некоторые окончательные результаты.

Вероятность того, что в системе находится k требований:

$$p_k = \begin{cases} \frac{1}{k}\left(\frac{\lambda}{\mu}\right)^k p_0, & 1 \leqslant k \leqslant n, \\ \frac{1}{n!\,n^{k-n}}\left(\frac{\lambda}{\mu}\right)^k p_0, & n \leqslant k \leqslant n+m; \end{cases} \tag{2.50}$$

$$p_0 = \frac{1}{\sum\limits_{k=0}^{n-1} \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k + \frac{1}{n!\left(1-\frac{\lambda}{n\mu}\right)}\left(\frac{\lambda}{\mu}\right)^n\left[1-\left(\frac{\lambda}{n\mu}\right)^{m+1}\right]} \tag{2.51}$$

Вероятность того, что поступившее требование получит отказ

$$p_{\text{отк}} = \frac{1}{n!\,n^m}\left(\frac{\lambda}{\mu}\right)^{n+m} p_0. \tag{2.52}$$

Средняя длина очереди

$$L_{\text{оч}} = \frac{p_n}{\left(1-\frac{\lambda}{n\mu}\right)^2}\left[\frac{\lambda}{\mu} - (m+1)\left(\frac{\lambda}{n\mu}\right)^{m+1} + m\left(\frac{\lambda}{n\mu}\right)^{m+2}\right]. \quad (2.53)$$

Среднее число требований в системе

$$L_{\text{сист}} = L_{\text{оч}} + \frac{1-\left(\frac{\lambda}{n\mu}\right)^{m+1}}{1-\frac{\lambda}{n\mu}} np_n + p_0 \sum_{k=1}^{n-1} \frac{1}{(k-1)!}\left(\frac{\lambda}{\mu}\right)^k. \quad (2.54)$$

— Теперь стоит рассмотреть пример 3 (продолжение примера 2). Поработав некоторое время по методу, описанному в примере 2, инспектора почувствовали, что такой образ действия довольно утомителен, и не позволяет проводить внимательную проверку. Поэтому они решили проверять путевые листы так, как они это делали в примере 1, — тратя в среднем по 5 мин на каждую проверку. Но чтобы проверить как можно больше машин, они решили допускать образование очереди. Если очередь ничем не ограничена, то, как было сказано в описании примера 2, она очень скоро начнет расти и движение застопорится. Поэтому инспектора решили, что очередь не должна превышать 3 машин. Если проезжающий грузовик застает в очереди 3 машины, он не останавливается.

Определим вначале p_0 — вероятность того, что все инспектора свободны. По формуле (2.51):

$$p_0 = \frac{1}{\sum_{k=0}^{2} \frac{1}{k!}\cdot 5^k + \frac{1}{3!\left(1-\frac{5}{3}\right)}\cdot 5^3\left[1-\left(\frac{5}{3}\right)^4\right]} \approx \frac{1}{231{,}5}$$

Вероятность того, что проезжающий грузовик не будет остановлен

$$p_{\text{отк}} = \frac{5^6}{3!3^3} p_0 \approx 0{,}42.$$

Напомним, что в примере 1, который отличается от примера 3 только тем, что число мест в очереди $m=0$, вероятность отказа была равна 0,53.

Определим среднюю длину очереди. Для этого найдем p_n — вероятность того, что около пункта ГАИ находятся точно 3 грузовика

$$p_n = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n p_0 \approx 0{,}09,$$

а затем по формуле (2.53) найдем

$$L_{\text{оч}} = \frac{0{,}09}{\left(1-\frac{5}{3}\right)^2}\left[\frac{5}{3} - 4\left(\frac{5}{3}\right)^4 + 3\left(\frac{5}{3}\right)^5\right] \approx 1{,}9.$$

— То, что я вам рассказал, — продолжал Александр Андреевич после некоторого молчания, — это, так сказать, азы теории массового обслуживания, самые простые вещи. Как и во всякой математической теории, в массовом обслуживании есть свои «начала» и свой «передний край» с множеством нерешенных проблем. Мы с вами только постучались в эту науку, ну, может быть, чуть-чуть приоткрыли дверь.

— А нельзя ли нам войти? Ну хотя бы в прихожую, — спросил Борис.

— В прихожую? — рассмеялся Александр Андреевич. — Нет, боюсь, что не получится — для этого нам понадобился бы сложный математический аппарат, и сложность эта возрастала бы с каждым шагом. Да и что

можно увидеть из прихожей? Разве только то, куда следовало бы идти дальше. Впрочем, — продолжал он после минутного размышления, — я могу вас ввести в прихожую и даже — более того — провести по комнатам, но только в качестве посетителей, осматривающих сложное производство.

— То есть вы предлагаете совершить экскурсию по теории массового обслуживания? — спросил Стрелкин.

— Что-то вроде этого, небольшой обзор, если вы еще не устали, — Александр Андреевич вопросительно посмотрел на спутников.

— И не будет формул? — полувопросительно-полуутвердительно сказал Стрелкин.

— Ну, может быть, самые простые.

— Мы не устали, Александр Андреевич, — сказала Галя, — и с интересом вас слушаем.

— Ну что ж, тогда давайте начнем. Итак, мы сделали много предположений относительно структуры и вероятностных характеристик некоторых систем, наложили разные ограничения и для самых простых случаев сделали расчеты. Вот, например, одноканальная система с неограниченной очередью. Возьмем ее за основу. Мы рассчитали эту систему при условии, что в нее поступает простейший поток, а обслуживание производится по показательному закону.

— А что делать в остальных случаях? — поинтересовался Стрелкин.

— Для одноканальных систем с неограниченной очередью и простейшим потоком существует выражение (формула Полячека—Хинчина) для среднего числа требований в системе, которое справедливо при любом распределении продолжительности обслуживания

$$L_{\text{сист}} = \rho + \frac{\rho^2 + \lambda^2 \sigma^2 \mu}{2(1-\rho)}. \qquad (2.55)$$

Здесь σ_μ^2 — дисперсия продолжительности обслуживания. Если обслуживание распределено по показательному закону, то, как мы видели, $\sigma_\mu^2 = 1/\mu^2$ и (2.55) превращается в (2.42). Из (2.55) получим выражение для $L_{оч}$, $W_{сист}$ и $W_{оч}$ в случае одноканальной системы с пуассоновским входящим потоком:

$$L_{оч} = \frac{\rho^2 + \lambda^2 \sigma_\mu^2}{2(1-\rho)}; \tag{2.56}$$

$$W_{сист} = \frac{1}{\mu} + \frac{\rho/\mu + \lambda\sigma_\mu^2}{2(1-\rho)}; \tag{2.57}$$

$$W_{оч} = \frac{\rho/\mu + \lambda\sigma_\mu^2}{2(1-\rho)}. \tag{2.58}$$

— Интересно посмотреть, какой получится по этой формуле простой составов в ожидании технического осмотра в парке прибытия? — спросил Стрелкин.

— Для этого нужна ваша помощь, — ответил Александр Андреевич. — Приведите, пожалуйста, исходные данные.

— Пусть в этот парк прибывают 70 поездов с трех направлений и распределение интервалов прибытия близко к показательному. Среднее время осмотра одного состава равно 12 мин. В парке работает одна бригада осмотрщиков.

— Коэффициент вариации длительности осмотра $v\mu = 0{,}3$, — добавил Борис.

— При этих данных, — сказал Александр Андреевич, — получим:

интенсивность входящего потока

$$\lambda = \frac{70}{24} = 2{,}91 \text{ поездов/ч};$$

интенсивность обслуживания

$$\mu = \frac{60}{12} = 5 \text{ поездов/ч};$$

дисперсия длительности обслуживания

$$\sigma_\mu^2 = \frac{\nu_\mu^2}{\mu^2} = \frac{0{,}3^2}{5^2} = 0{,}004 \text{ ч}^2;$$

нагрузка бригады

$$\rho = \frac{\lambda}{\mu} = \frac{2{,}91}{5} \approx 0{,}6 \text{ и}$$

среднее время ожидания по формуле (2.58):

$$W_{\text{оч}} = \frac{\frac{0{,}6}{5} + 2{,}91 \cdot 0{,}004}{2(1-0{,}6)} = 0{,}16 \text{ ч}.$$

— Это совпадает, пожалуй, с реальным простоем вагонов на крупных сортировочных станциях, — подтвердил Стрелкин.

— Таким образом, формула Полячека — Хинчина позволила снять одно ограничение, а именно: условие о показательном распределении времени обслуживания, — заключил Александр Андреевич. — Какое условие естественно снять теперь?

— Условие о простейшем характере входящего потока, — ответил Борис, — пусть теперь и входящий поток произвольный.

— Произвольный рекуррентный, — уточнил Александр Андреевич, — т. е. стационарный ординарный поток с ограниченным последействием и произвольным законом распределения интервалов между моментами

поступления требований в систему. Для системы с таким входящим потоком точных формул, к сожалению, нет, однако есть оценки, т. е. верхняя и нижняя границы для среднего числа требований в очереди

$$L_{\text{оч}} \leqslant \frac{\lambda^2\left(\sigma_\lambda^2+\sigma_\mu^2\right)}{2(1-\rho)}, \tag{2.59}$$

здесь σ_λ^2 и σ_μ^2 — дисперсии интервалов входящего потока и обслуживания; λ и ρ — интенсивность входящего потока и нагрузка системы.

$$L_{\text{оч}} \geqslant \frac{\lambda^2\sigma_\mu^2}{2(1-\rho)} - \frac{\rho}{2}. \tag{2.60}$$

Если на входящий поток наложить некоторые ограничения, можно использовать более точные оценки. Например, если входящий поток эрланговский,то

$$L_{\text{оч}} \leqslant \frac{\rho^2\left(\nu_\lambda^2+\nu_\mu^2\right)}{2(1-\rho)} + \frac{1-\rho}{2}\left(1-\nu_\lambda^2\right); \tag{2.61}$$

$$L_{\text{оч}} \geqslant \frac{\rho^2\left(\nu_\lambda^2+\nu_\mu^2\right)}{2(1-\rho)} - \frac{\rho}{2}\left(1-\nu_\lambda^2\right). \tag{2.62}$$

Здесь $\nu_\lambda = \lambda\sigma_\lambda$ и $\nu_\mu = \mu\sigma_\mu$ —коэффициенты вариации интервалов входящего потока и обслуживания. Если входящий поток простейший, то $\nu_\lambda = 1$ и формулы (2.61) и (2.62) превращаются в формулу Полячека—Хинчина.

— Заметим, что величина

$$\frac{\rho^2\left(\nu_\lambda^2+\nu_\mu^2\right)}{2(1-\rho)}$$

находится между верхней и нижней оценками и, следовательно, может быть использована как приближенное значение $L_{оч}$. Если воспользоваться формулой Литла (2.32), то можно получить приближенную формулу и для среднего времени ожидания в очереди

$$W_{оч} = \frac{\rho^2 \left(\nu_\lambda^2 + \nu_\mu^2 \right)}{2\lambda (1 - \rho)}. \qquad (2.63)$$

— Меня интересует, как изменится простой составов, если распределение интервалов их прибытия будет не показательным, как в предыдущем примере. Пусть коэффициент вариации входящего потока будет равен, например, 0,7, — сказал Стрелкин.

— Тогда среднее время ожидания по формуле (2.63) составит

$$W_{сч} = \frac{0{,}6^2 (0{,}7^2 + 0{,}3^2)}{2 \cdot 2{,}91 (1 - 0{,}6)} = 0{,}09 \text{ ч}$$

и по сравнению с предыдущим примером, когда интервалы прибытия имели показательное распределение, оно уменьшилось на 0,16—0,09=0,07 ч, или примерно на 40%.

— Это довольно-таки существенно,—заметил Стрелкин.

— А если входящий поток даже не реккурентный? — спросила Галя.

— Давайте разберемся, что же будет в таком случае. Еще раз напомним: если поток не рекуррентный, то нарушено по крайней мере одно из условий: ординарность, стационарность, ограниченность последействия. Если поток неординарен, а остальные условия выполнены, ничего страшного нет: есть формулы, позволяющие во многих случаях свести системы с неординарным входящим

потоком к системам с ординарным потоком. Если поток нестационарен, все гораздо сложнее: такие системы практически нельзя рассчитать аналитическими методами. Кое-что удается сделать, если входящий поток имеет периодически изменяющуюся интенсивность. Например, известно, что предельные вероятности состояний такой системы — периодические функции, причем их период совпадает с периодом входящего потока. Совсем плохо дело обстоит, если поток имеет неограниченное последействие, т. е. его интервалы связаны корреляционной зависимостью. Тут математики пока мало что смогли сделать.

— Все системы, которые мы рассчитывали, — продолжал Александр Андреевич, — имели простую дисциплину обслуживания: первым пришел — первым обслужен. Но, может быть, вы вспомните: в начале нашего разговора я называл и другие дисциплины.

— Вы говорили об инверсионном порядке обслуживания: последний пришел — первым обслужен; случайном порядке: выбор из очереди на обслуживание производится случайным образом и о приоритетном обслуживании.

— Совершенно верно, Боря. Вот на последнем порядке выбора на обслуживание — приоритетном — я хочу остановиться подробнее. Исследования по приоритетным системам составляют значительную часть всех работ в области теории массового обслуживания. И это не удивительно, приоритетные системы встречаются очень часто. Например, если со станции должно быть отправлено несколько поездов, то высший приоритет имеют пассажирские поезда. Если на аэродроме одна взлетно-посадочная полоса, то высший приоритет имеют самолеты, заходящие на посадку. На железнодорожном переезде поезда имеют более высокий приоритет при обслуживании, чем автомобили.

Общая постановка задачи такова: в систему массового обслуживания поступает несколько потоков требований. Требования из разных потоков в общем случае имеют разные распределения длительности обслуживания. Все требования, принадлежащие одному потоку, имеют одинаковый приоритет, требования из разных потоков имеют разные приоритеты. Все приоритеты пронумерованы. Требования, имеющие самый высокий приоритет (самому высокому приоритету может быть присвоен, например, первый номер), обслуживаются в первую очередь. Если требований первого приоритета в системе нет, начинают обслуживаться требования второго, более низкого приоритета и так далее. В последнюю очередь обслуживаются требования самого низкого приоритета. Возможны случаи, когда требование высокого приоритета, пришедшее в систему после начала обслуживания требования низкого приоритета, ждет, пока прибор не освободится.

— Но есть и другие примеры, — вставил Стрелкин. — Ремонтная бригада ведет профилактические работы на каком-то километре пути. В это время приходит сообщение о том, что в другом месте произошло повреждение пути. Ремонтная бригада будет немедленно переброшена туда, а профилактические работы на этом участке она закончит после устранения повреждения.

— В этом примере требование высокого приоритета не будет ожидать окончания обслуживания требования низкого приоритета, — сказал Борис.

— И тот и другой случай — примеры приоритетных систем, но приоритеты здесь разные, — сообщил Александр Андреевич.

— Значит, есть два разных вида приоритетов?

— Не два, а гораздо больше. Но эти два, пожалуй, самые важные и распространенные. В первом случае имеет место так называемый относительный прио-

ритет. Он характерен тем, что в системах с таким типом приоритета требование высокого приоритета, поступив в систему после начала обслуживания требования низкого приоритета, будет ждать, пока прибор не освободится. В примере, который сообщил нам уважаемый Стрелкин, мы познакомились с абсолютным приоритетом. В системах с абсолютным приоритетом высокоприоритетное требование, застав на приборе требование с низким приоритетом, «выталкивает» его и немедленно начинает обслуживаться.

— А что будет с «вытолкнутым» требованием?

— Здесь различают несколько случаев. Требование, обслуживание которого было прервано, впоследствии может вновь вернуться на прибор. Можно представить, что будучи «вытолкнутым», оно занимает место во главе очереди из требований с таким же приоритетом. В других случаях «вытолкнутое» требование может теряться.

— Если «вытолкнутое» ранее требование возвращается на прибор, время, уже затраченное на его обслуживание, учитывается? — спросил Борис.

— Это время учитывается в системах с дообслуживанием. Пример с ремонтной бригадой — это как раз система с абсолютным приоритетом и дообслуживанием. Но есть и другие системы, с обслуживанием заново.

— Как рассчитываются такие системы?

— Этот вопрос весьма сложен, и мне не удастся вам его подробно объяснить. Скажу только, что приоритетные системы рассчитывают, как правило, в предположении о том, что все входящие потоки простейшие. Еще можно отметить, что в системах с абсолютными приоритетами требования высшего приоритета ведут себя так, как будто никаких других требований в системе нет; поэтому характеристики для требования высшего приоритета — это то же самое, что характеристики для не-

приоритетной системы, в которую поступает только один поток требований.

Некоторые характеристики приоритетных систем в случае, когда требования всех потоков имеют одинаковое распределение длительности обслуживания, совпадают с аналогичными характеристиками неприоритетных систем. Например, среднее число всех требований в системе для систем с относительным приоритетом и системы с абсолютным приоритетом с дообслуживанием такое же, как и в неприоритетной системе с входящим потоком, равным суммарному входящему потоку приоритетной системы. То же самое, кстати, относится и к системам с инверсионным и случайным порядком выбора на обслуживание.

— Александр Андреевич, вы говорили, что кроме абсолютного и относительного бывают и другие приоритеты. Расскажите о них, пожалуйста, — попросила Галя.

— Действительно, есть и другие приоритеты. Например, чередующийся. Пусть в систему поступают два потока требований A и B. Если в данный момент на приборе находится требование из потока A, то все требования из этого потока имеют высший приоритет. Но вот в какой-то момент обслужено последнее из находящихся в системе требований потока A, и высший приоритет переходит к потоку B, если требования из этого потока присутствуют в системе. Теперь пришедшие требования потока A будут ждать до тех пор, пока в системе не останется ни одного требования потока B.

— Это какой-то очень сложный случай, — сказал Стрелкин, — такие системы, наверно, в жизни не бывают.

— Еще как бывают! — возразил Борис. — Вспомните флюорографию. Вы пришли — около кабинета сидят женщины. «Там мужчина?» — «Мужчина». Вы заходи-

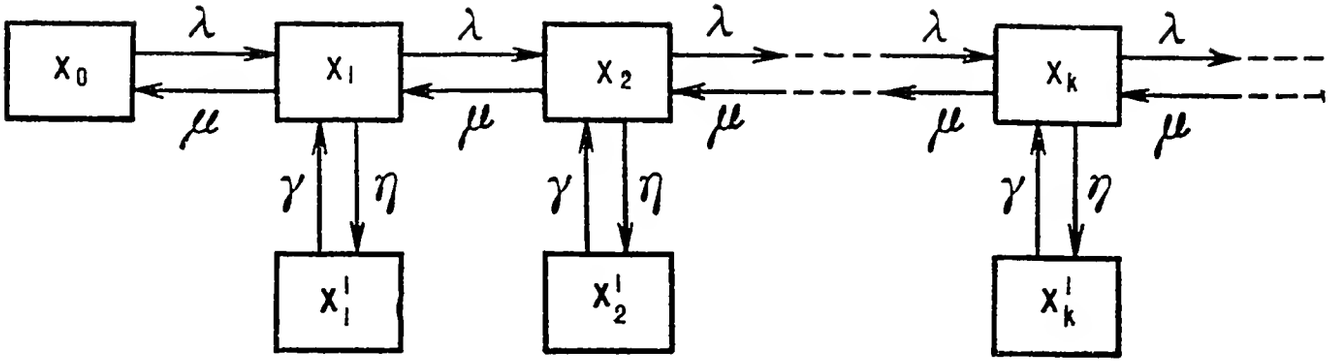


Рис. 2.14

те — у вас сейчас высший приоритет. Но если вы замешкались в раздевалке, пришли к кабинету, когда тот мужчина вышел, вы прождете, пока женщины, которые сидят в очереди, и которые еще придут потом, не пройдут флюорографию. Приоритет у вас будет низшим.

— Есть и другие приоритеты: динамические, смешанные и т. д., но, кажется, ясно, что типов приоритетов довольно много, и много типов можно еще выдумать. Наиболее часто на практике все же встречаются два типа: абсолютный и относительный.

— До сих пор мы говорили о системах с абсолютно надежным прибором, — продолжал Александр Андреевич. — Мы считали прибор всегда исправным и готовым к обслуживанию. Между тем, во многих практических задачах приходится учитывать возможность выхода прибора (или приборов, если их несколько) из строя. Это так называемые системы с ненадежным прибором. Давайте для простоты будем говорить только об одноканальных системах. В таких системах прибор периодически может выходить из строя. Происходит это в случайные моменты времени. После выхода прибора из строя начинается его ремонт. Продолжительность ремонта — также случайная величина.

Такую систему можно рассматривать как систему с двумя потоками требований и абсолютным приоритетом. Поток требований с высшим приоритетом — это

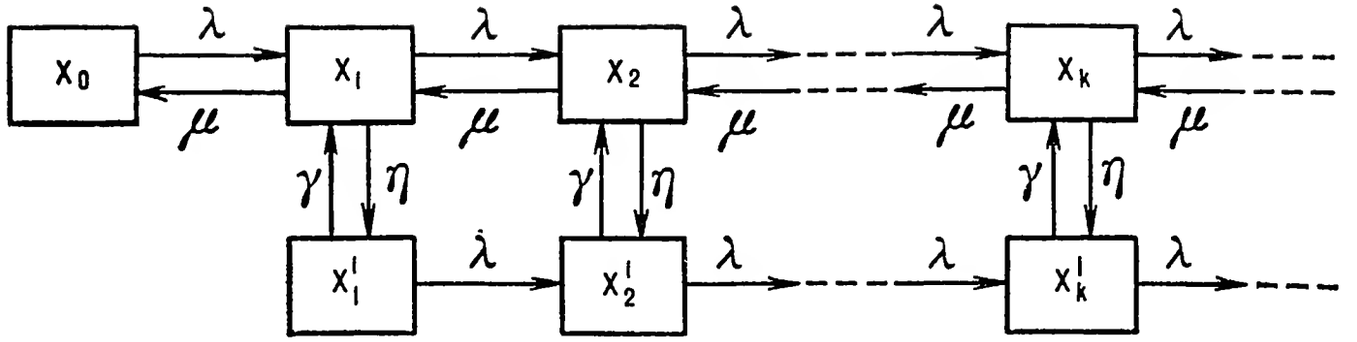


Рис. 2.15

поток неисправностей. В момент прихода такого «требования» обслуживание требования с низким приоритетом мгновенно прерывается и начинается «обслуживание» пришедшего высокоприоритетного «требования» — ремонт прибора. Некоторое отличие от «обычных» приоритетных систем состоит в том, что в системе с ненадежным прибором поломка происходит в периоды исправной работы, а в приоритетной системе требование высшего приоритета может прийти в момент, когда на приборе находится другое требование того же приоритета.

Давайте рассмотрим несколько примеров одноканальных систем, отличающихся друг от друга типами неисправностей. Будем считать, что прибор может выходить из строя только тогда, когда он обслуживает требования (т. е. выход из строя в периоды простоя невозможен). Рассмотрим случай, когда выход из строя ведет к полному прекращению работы системы как по обслуживанию, так и по приему требований. Пусть λ и μ означают интенсивности поступления и обслуживания требований, а η и γ — интенсивности выхода прибора из строя и его восстановления (ремонта). Граф состояний такой системы изображен на рис. 2.14. (Отметим, что если допустить возможность выхода прибора из строя в состоянии простоя, то добавится состояние x_0^1 с соответствующими связями с состоянием x_0.)

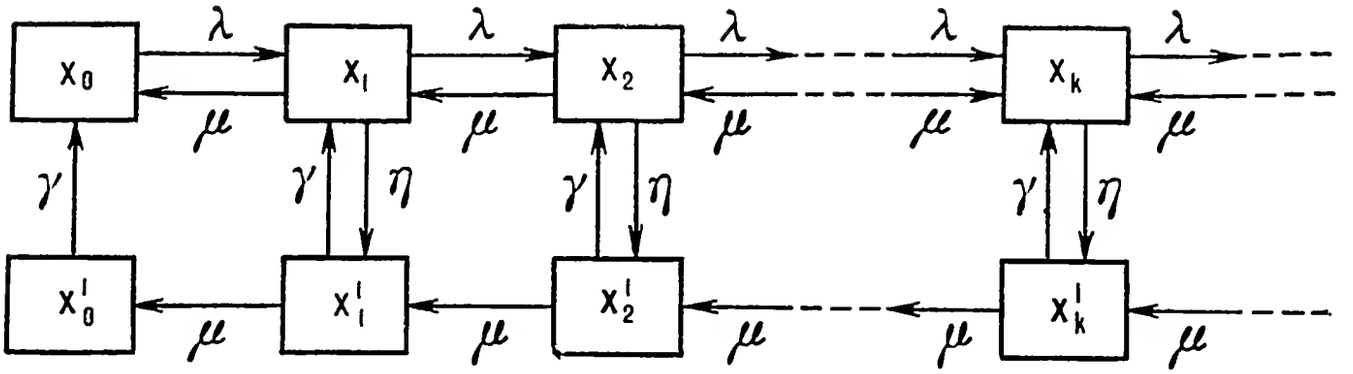


Рис. 2.16

Пусть во время выхода прибора из строя поступление новых требований в систему продолжается. Тогда граф состояний этой системы имеет вид, изображенный на рис. 2.15.

Возможен случай, когда поломка прибора приводит к тому, что прекращается прием новых требований, но производится обслуживание тех требований, которые имелись в системе к моменту поломки. Граф состояний этой системы изображен на рис. 2.16.

И, наконец, рассмотрим случай, когда во время поломки прибора происходит и прием новых требований, и обслуживание, но только обслуживаются требования в два раза медленнее, чем при исправном приборе. Граф состояний системы показан на рис. 2.17.

— Раз уж мы заговорили о свойствах самого прибора, то уместно упомянуть и о так называемых «системах с разогревом», — сказал Александр Андреевич. — Я поясню понятие «разогрев» на примере самой простой одноканальной системы с неограниченной очередью. Представьте себе такой прибор, который выключается после завершения обслуживания последнего из имеющихся в системе требований. Весь период простоя (т. е. период отсутствия требований в системе) этот прибор выключен. В момент поступления нового требования прибор вновь включается. Но вследствие того, что прибор находится в «холодном» состоянии, он не может

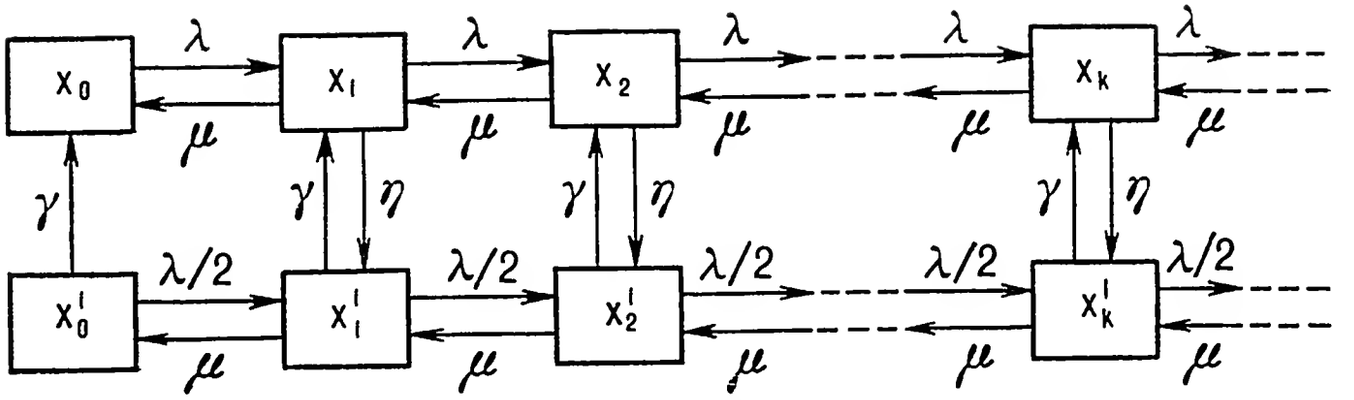


Рис. 2.17

начать обслуживать это требование немедленно, а должен предварительно «разогреться». После окончания периода разогрева прибор готов к обработке скопившихся к этому моменту требований. Продолжительность разогрева может быть как постоянной величиной, так и случайной — с некоторой функцией распределения продолжительности разогрева.

Давайте такую систему изобразим графически (рис. 2.18), причем α будет обозначать интенсивность разогрева, т. е. величину, обратную среднему времени разогрева, а символами x_k^p обозначим состояния разогрева, где k означает количество требований, собравшихся в системе в процессе разогрева.

— Наверно, на практике встречаются не только системы с разогревом, но и системы с «остыванием». По-видимому, в теории массового обслуживания есть и такой термин, — предположил Борис.

— Такой термин есть, и системы с остыванием в теории массового обслуживания рассматриваются. Различают даже несколько их типов. Отличаются они по поведению прибора в моменты, когда началось остывание (оно начинается, когда система закончила обслуживание последнего из имевшихся в ней требований) и пришло новое требование. Здесь для разных систем возможны разные случаи: период остывания доходит до

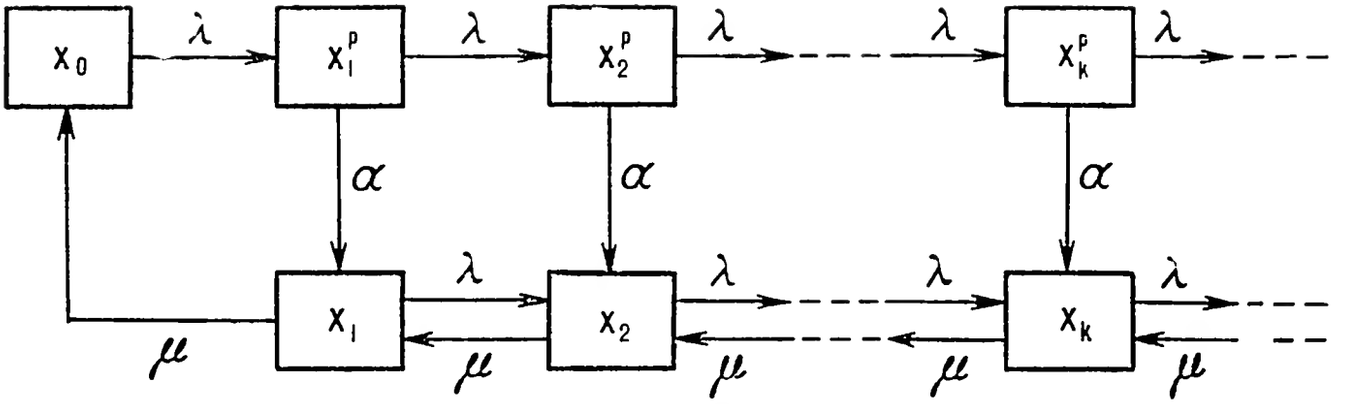


Рис. 2.18

конца, а затем сразу начинается период разогрева; остывание немедленно прекращается, начинается разогрев, причем он ничем не отличается от разогрева, который начинается в условиях «холодной» системы; так же, как и в предыдущем случае, сразу начинается разогрев, причем система только «доразогревается» с учетом того, насколько она успела остыть; прибор считается разогретым в течение всего периода остывания, поэтому обслуживание требования начинается сразу после прибытия. Так же, как и в случае системы с разогревом, время остывания может быть как случайным, так и детерминированным.

— Как много, оказывается, существует систем массового обслуживания! — воскликнула Галя.

— И, наверно, то, что рассказал нам Александр Андреевич, — это еще не все описанные и изученные системы, — продолжил Борис.

— Конечно, нет! — подтвердил Александр Андреевич. —Вот вам еще один тип системы массового обслуживания. Мы уже говорили о системах с неординарным входящим потоком, когда требования поступают группами, численности которых случайны. Точно так же можно рассматривать и групповое обслуживание. Правда, здесь дело обстоит несколько сложнее и имеется больше частных случаев, чем в системах с неординарным потоком.

Часто рассматривают системы, в которых величина обслуживаемой группы не превосходит величины k (например, лифт). Если к моменту начала обслуживания новой группы в очереди больше, чем k требований, часть требований остается в очереди в ожидании следующих циклов обслуживания. В некоторых системах запрещается брать на обслуживание меньше, чем k требований, и тогда прибор может простаивать в ожидании, пока длина очереди достигнет k. Естественно, что могут существовать системы, в которых оба эти условия совмещены (часто так бывает с маршрутными такси). Величина k может быть как детерминированной, так и случайной.

Говоря о групповом обслуживании, следует сказать и о системах с групповым обслуживанием по расписанию. В таких системах обслуживание начинается в определенные моменты времени, не зависящие от величины очереди.

— Для таких систем можно привести много примеров из области транспорта, — сказал Стрелкин. — Пригородные и дальние пассажирские поезда, пассажирские теплоходы, самолеты.

— Городской транспорт: автобусы, трамваи, троллейбусы, метро, — добавила Галя.

— А что, если в момент начала обслуживания в очереди нет ни одного требования? — поинтересовался Борис.

— Тогда будет обслуживаться группа из 0 требований. Так часто бывает на подвесной канатной дороге. Подошла очередная кабина (или кресло), ее никто не занял, и она уходит пустой.

— Мы сейчас говорили об одноканальных системах массового обслуживания, — продолжил свой рассказ Александр Андреевич. — Все те особенности, о которых мы упомянули: приоритеты, ненадежность, разогрев и

т. д., могут быть свойственны и многоканальным системам, состоящим из нескольких приборов. Но многоканальным системам присущи и некоторые особенности, которых не может быть у одноканальных систем, и поэтому стоит поговорить о них отдельно.

Прежде всего, можно отметить, что приборы могут быть разными, т. е. обслуживать требования в соответствии с разными законами распределения времени обслуживания. Такие системы — довольно сложные для исследования объекты, поэтому давайте сразу договоримся рассматривать только системы с одинаковыми приборами.

Вначале рассмотрим обычную многоканальную систему с неограниченной очередью. В случае когда на вход этой системы поступает простейший входящий поток, а продолжительность обслуживания распределена по показательному закону, как мы уже видели, существуют формулы, по которым систему можно рассчитать. При других предположениях относительно входящего потока и обслуживания точных формул для среднего времени ожидания и средней длины очереди нет. Для таких систем известна лишь формула для среднего числа занятых приборов в произвольный момент времени.

Если в системе n приборов, λ — интенсивность потока, μ — интенсивность работы одного прибора, $\rho = \lambda/n\mu$ — нагрузка, то среднее число работающих приборов в системе M можно подсчитать так:

$$M = n\rho.$$

Хотя точных формул для расчета основных характеристик многоканальных систем с произвольными распределениями нет, существуют оценки, позволяющие найти их приближенные значения. Для этого подбирают две

одноканальные системы так, что характеристики в одной из них больше, чем в многоканальной системе, а в другой меньше. А для этих одноканальных систем есть свои оценки, например (2.59)—(2.62).

— А как подбираются эти системы? — поинтересовался Борис. — Это делается каким-то очень сложным способом?

— Это делается весьма просто. Как подобрать одноканальную систему, в которой среднее время ожидания больше, чем в многоканальной? Вспомним, как работает многоканальная система. Все требования, поступившие в систему и заставшие при этом все приборы занятыми, становятся в общую очередь. Как только освобождается какой-либо прибор, первое стоящее в очереди требование немедленно его занимает. Теперь, представьте себе, мы вводим в этой системе другой порядок. Каждое поступающее в систему требование уже заранее знает, какой прибор должен его обслуживать. Задается это таким правилом: первое поступившее требование идет к первому прибору, второе — ко второму, n-е к n-му прибору (система n-канальная), затем $(n+1)$-е требование идет к первому прибору и т. д. Имеет место определенный цикл, и такой порядок обслуживания называется циклическим. Таким образом, 1-е, $(n+1)$-е, $(2n+1)$-е и т. д. требования становятся в очередь к первому прибору, 2-е, $(n+2)$-е, $(2n+2)$-е и т. д. — в очередь ко второму прибору и т. д. В системе уже не одна очередь, а n. При таком порядке обслуживания возможны, например, случаи, когда первый прибор свободен и простаивает, а ко второму прибору стоит очередь. Если в первоначальной многоканальной системе приборы были взаимозаменяемы и работали «сообща», то в системе с циклическим порядком обслуживания каждый прибор работает только «за себя». Совершенно очевидно, что при такой оптимальной

организации обслуживания среднее число требований и их среднее время ожидания в очереди больше, чем в обычной многоканальной системе.

Теперь давайте внимательно присмотримся к системам с циклическим порядком обслуживания. Как вы уже, наверно, обратили внимание, многоканальная система фактически состоит из n одноканальных систем.

Прибор в каждой одноканальной системе тот же, что и в многоканальной, а входящий поток имеет интенсивность в n раз меньшую и получается из входящего потока многоканальной системы путем выборки из него каждого n-го требования. Например, если первоначальный поток был простейшим, то полученный таким образом поток будет эрланговским n-го порядка. Среднее время ожидания в такой одноканальной системе, конечно, то же самое, что и в многоканальной циклической системе, а следовательно, больше, чем в исходной многоканальной системе. А для одноканальной системы можно уже использовать верхнюю оценку (2.59). Если же входящий поток был простейшим, в оценочной одноканальной системе он эрланговский и можно использовать верхнюю оценку (2.61).

— Видимо, похожим образом подбирается одноканальная система и для получения нижней оценки? — спросил Борис.

— Для получения нижней оценки делается наоборот: берется одноканальная система с точно таким же входящим потоком, что и в многоканальной, но с обслуживанием в n раз более быстрым. Причем, если в первом случае поток «растягивался» за счет n-кратного разряжения, то в этом случае время обслуживания некоторого требования ровно в n раз короче обслуживания точно такого же по счету требования, поступившего в многоканальную систему.

Для получения нижней оценки удобно проводить рассуждения не для среднего времени ожидания требования в очереди, как в предыдущем случае, а для среднего времени его нахождения в системе (т. е. и в очереди, и на обслуживании). Но, как мы уже видели на примере одноканальных систем, эти две величины отличаются друг от друга на среднее время обслуживания, а оно равно $1/\mu$.

Итак, входящий поток в одноканальной и многоканальной системах одинаковый, в одноканальной системе прибор работает в n раз быстрее, чем один прибор в n-канальной. Таким образом, если в многоканальной системе работают все n приборов, скорости обслуживания у них равны. Но в многоканальной системе могут быть такие случаи, когда в системе есть требования, а она работает не с максимальной скоростью. Например, в системе находится всего одно требование, один прибор занят, остальные $(n-1)$ простаивают, и общая скорость обслуживания в n раз меньше максимальной. В одноканальной системе такого быть не может: пока остается хотя бы одно требование, система работает с максимальной скоростью. Это объясняет, почему в рассматриваемой одноканальной системе среднее время нахождения требований в системе меньше, чем в исходной многоканальной системе. Использование этого факта, а также нижней оценки (2.60) для одноканальной системы (или нижней оценки (2.62), если входящий поток — эрланговский) позволяет получить нижние оценки для характеристик многоканальной системы.

Говоря о многоканальных системах, следует упомянуть о порядке поступления требований на обслуживание. О циклическом порядке мы только что говорили. Как мы видели, такая организация — не самая лучшая с точки зрения поступающих требований, однако такие системы довольно часто встречаются.

— Я могу привести пример, — предложил Стрелкин. — Кассы предварительной продажи билетов обычно находятся в отдельном помещении. Несколько касс обслуживают одно направление. Пассажиры заходят в это помещение и сразу становятся в очередь к какой-то кассе. Конечно, здесь нет четкого циклического деления по кассам, но приблизительно так и получается.

— Но в вашем примере не бывает так, что одна касса свободна, а к другой стоит очередь, как случается в циклических системах, — возразила Галя.

— Конечно, как только одна очередь заметно укоротится, кто-нибудь из длинной очереди в нее перейдет, — согласился Стрелкин.

— Здесь мы столкнулись с примером более сложной системы, — объяснил Александр Андреевич. — Порядок присоединения к очереди в первом приближении можно считать циклическим, но, кроме этого, мы сталкиваемся здесь с таким явлением, как возможность перехода из одной очереди в другую. Такими системами, по сути дела, является большинство магазинов, мастерских и других учреждений, где в очереди стоят люди. Как только кто-то видит, что в соседней очереди он может простоять меньше, он постарается в нее перейти.

— Раз системы с возможностью перехода из очереди в очередь существуют в жизни, наверно, они изучаются и в теории массового обслуживания? — предположил Борис.

— Изучаются, — подтвердил Александр Андреевич,— но вряд ли я смогу сообщить вам какие-нибудь простые результаты, касающиеся этих систем. Я хотел бы вернуться к разговору о порядке доступа требований непосредственно на приборы и поговорить о еще одном типе систем. Как вы помните, в системах с циклическим обслуживанием требованиям приходится ожидать в среднем больше, чем в обычных системах. Легко за-

метить, что среднее время работы приборов в этой системе не отличается от той же характеристики в обычной системе. А вот пример системы, в которой требования ожидают столько же, сколько в простых многоканальных системах, а приборы работают по-другому. Представьте себе, что все приборы пронумерованы от 1 до n и, если свободно несколько приборов, требование занимает прибор с меньшим номером. В такой системе максимальную нагрузку будет испытывать первый прибор, а минимальную — n-й.

— По-видимому, примером может служить двухканальная система, где один прибор основной, а второй — резервный, и подключается, когда возникает очередь, — сказал Борис.

— Еще один класс многоканальных систем, о которых я хотел рассказать, — продолжил рассказ Александр Андреевич, — это системы с взаимопомощью между приборами. При наличии очереди они работают так же, как обычные многоканальные системы. Но когда количество требований в системе становится меньше числа приборов, освободившиеся приборы «помогают» тем, которые еще заняты обслуживанием требований. Например, если в двухканальной системе один прибор освободился, он приходит на помощь работающему прибору и вместе они заканчивают обслуживание требования в два раза быстрее. Если в процессе такой «совместной» работы поступит новое требование, «прибор-помощник» переключается на его обслуживание и каждый прибор обслуживает свое требование.

— Александр Андреевич, вы говорили, что приборы могут располагаться не только параллельно, но и последовательно. Кажется, вы называли такие системы многофазовыми? — спросила Галя.

— Да, действительно, я говорил о многофазовых системах. Помните, я упоминал о том, что многофазо-

вые системы могут быть одновременно и многоканальными на всех или нескольких фазах. Трудности при исследовании многофазовых систем возникают в основном из-за того, что потоки требований, переходящих с одной фазы на другую, как правило, имеют весьма сложную структуру. Давайте будем пояснять наши рассуждения на примере двухфазовой системы. Так вот, поток, выходящий с первой фазы, как правило, имеет неограниченное последействие. Этот поток поступает на вторую фазу, и мы, таким образом, сталкиваемся с системой массового обслуживания, на вход которой поступает поток с неограниченным последействием. Совсем недавно мы говорили о том, что такие системы почти не поддаются аналитическому исследованию. Правда, есть небольшое число исключений, когда выходящий поток все же является рекуррентным. К счастью, к ним относится (может быть, вы помните, мы упоминали об этом) такой важный для нас случай, когда входящий поток первой фазы — простейший, а обслуживание на всех приборах первой фазы — показательное. Более того, в этом случае выходящий поток так же, как и входящий, — простейший. Как правило, когда возникает необходимость рассчитать многофазовую систему, делают предположение о том, что входящий поток является простейшим, а продолжительность обслуживания распределена по показательному закону.

Однако трудности при расчете многофазовых систем возникают не только из-за свойств выходящего потока, но и из-за ограничений, часто налагаемых на порядок прохождения требований через фазы. Самый простой случай — это когда очередь перед второй фазой не ограничена. Если и перед первой фазой очередь не ограничена, то система просто распадается на две однофазовые системы с неограниченными очередями. Но очень часто число мест для ожидания на второй фазе

ограничивается, как это было в примере с сортировочной станцией, который мы разбирали в начале нашего разговора. И вот тут возникают различные частные случаи. Что произойдет, если вторая очередь заполнит все отведенные ей места, и первая фаза закончит обслуживание еще одного требования?

— Тогда прибор, закончивший обслуживание этого требования, заблокируется. Так было в примере со станцией, — ответил Стрелкин.

— Это только один из возможных случаев. В этой системе имеется обратная связь между второй и первой фазами: состояние первой фазы зависит от того, что делается на второй. Бывают системы с ограниченной очередью второй фазы, но без обратной связи. В них требование, обслуженное на первой фазе и заставившее вторую фазу занятой, теряется, уходит из системы недообслуженным.

Частным случаем являются системы с отсутствием мест для очереди на всех фазах, кроме первой. Среди них могут быть системы с блокировкой и с потерями между фазами, но возможен и еще один случай, когда при прохождении одним требованием всех фаз остальные требования ожидают в очереди перед первой фазой, пока обслуживаемое требование не покинет систему. Здесь тоже происходит блокировка, но блокируются все фазы системы, кроме той, которая занята обслуживанием.

— Итак, вы видите, — подытожил Александр Андреевич, — сколь разнообразны системы массового обслуживания. Системы, которые встречаются в практической деятельности человека, могут быть одновременно многофазовыми, многоканальными, с приоритетным обслуживанием, ограниченными очередями, блокировкой, сложными распределениями интервалов входящего потока и обслуживания. Конечно, если вам когда-ни-

будь придется работать с такой сложной системой, для ее расчета вам скорее всего не хватит того математического аппарата, о котором мне удалось рассказать, однако вы сможете правильно описать систему и поставить задачу ее решения.

— Александр Андреевич, — сказал Борис, — я давно хочу задать вам один вопрос. Во всех случаях, о которых вы рассказывали, требования приходили в систему из какого-то неизвестного нам источника, и моменты их поступления никак не зависели от состояний системы. Всегда мы рассуждаем, например, так: «Пусть в систему массового обслуживания поступает простейший поток требований...». И поток будет одним и тем же и когда в системе находится одно требование, и когда их будет десять тысяч. Между тем, я могу привести случай, когда поток очень сильно зависит от того, сколько требований уже находится в системе, а при некоторых ее состояниях вообще может прекращаться. Моим примером будет работа такой системы обслуживания, как локомотивное депо. Я не очень хорошо представляю себе локомотивное хозяйство и, наверно, допускаю технологические неточности, но дело обстоит так. К локомотивному депо приписано несколько локомотивов, например 10. Время от времени все локомотивы проходят профилактику. Кроме того, с некоторыми происходят поломки, и они направляются в депо на ремонт. Чем больше локомотивов находится на профилактике и ремонте, тем меньше их может поступать в депо, и поток требований в систему уменьшается. В предельном случае, когда все 10 локомотивов находятся в депо, поток новых поступлений прекращается. Такая система не подходит ни под одно из определений, с которыми вы нас познакомили.

— Вы затронули очень интересную и весьма существенную область в теории массового обслуживания, —

ответил после некоторого размышления Александр Андреевич. — Действительно, Борин пример не укладывается в рамки той теории, о которой я рассказывал Для того чтобы объяснить, в чем тут дело, мне нужно ввести еще одну небольшую классификацию в теорию массового обслуживания. Все системы массового обслуживания дополнительно к тем классам, о которых мы уже говорили, можно разделить на две части: разомкнутые системы и замкнутые. Все без исключения системы, о которых мы говорили, являются разомкнутыми. Разомкнутость состоит в том, что между состоянием системы и входящим потоком нет обратной связи. Все требования как бы поступают из неисчерпаемого (бесконечного) источника в соответствии с каким-то вероятностным законом. Наряду с разомкнутыми существуют замкнутые системы, в которых эта обратная связь есть. Здесь требования поступают уже из некоторого ограниченного (конечного) источника. Общее количество требований в источнике и системе постоянно, и чем больше их в системе, тем меньше в источнике, поток же новых требований в систему уменьшается. В Борином примере система — это депо, где локомотивы ремонтируются или ожидают ремонта, а источник— та часть железнодорожной сети, на которой работают локомотивы, приписанные к этому депо.

К замкнутым системам относится все то, что мы говорили о разомкнутых: они могут быть многоканальными, многофазовыми, приоритетными и т. д. Вообще-то, если говорить строго, то практически все встречающиеся системы являются замкнутыми. Например, магазин мы считали системой разомкнутой, однако если принять во внимание, что население Земли — конечная величина и очередь в магазин не может этой величины превзойти, то надо считать магазин замкнутой системой массового обслуживания. Однако во многих случа-

ях источник требований столь велик, что его вполне можно считать бесконечным. Даже если ограничить количество возможных посетителей магазина числом жителей близлежащего района, то это число настолько превосходит реальное количество посетителей, одновременно находящихся в магазине, что такую систему практически без погрешностей можно считать разомкнутой. Однако в Борином примере число локомотивов невелико, и если считать депо разомкнутой системой массового обслуживания, то цифры, которые мы получим в результате расчета, могут очень сильно отличаться от того, что происходит в реальных условиях.

— При каком числе требований в системе и источнике замкнутую систему можно с достаточной точностью заменить разомкнутой? — спросила Галя.

— Это сложный вопрос. Число требований, циркулирующих между источником и системой, — не единственная и, может быть, даже не главная величина, влияющая на погрешность при замене замкнутой системы разомкнутой. Тут очень важно, какая часть всех требований находится в источнике, а это зависит от нагрузки системы. Если нагрузка велика и бóльшая часть требований систематически стоит в очереди, такая система плохо аппроксимируется разомкнутой моделью. В этой системе в источнике находится мало требований, и каждое изменение числа требований в нем может сильно влиять на поток поступлений в систему. Такой поток очень трудно считать рекуррентным. Наоборот, если нагрузка системы мала, бóльшая часть требований систематически находится в источнике, поток поступлений в систему имеет примерно постоянную интенсивность, и модель разомкнутой системы может оказаться вполне пригодной.

— А в чем смысл замены замкнутой системы разомкнутой? — поинтересовался Борис.

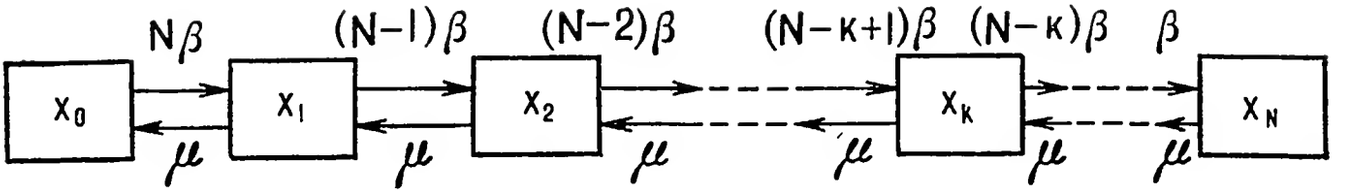


Рис. 2.19

— Дело в том, что замкнутые системы в общем случае рассчитывать значительно труднее, чем разомкнутые. И очень часто получить результат удается именно путем размыкания системы.

— Интересно, есть ли замкнутые системы, которые можно посчитать аналитически какими-то прямыми методами? — спросила Галя.

— Если есть, то это, наверно, системы, где все случайные величины распределены по показательным законам, — предположил Борис.

— Вы совершенно правы. Этот случай действительно аналитически рассчитывается, и расчет настолько прост, что, как мне кажется, о нем можно рассказать. Давайте будем говорить, как это мы делали и раньше, об одноканальной системе. В замкнутой системе важной величиной является период нахождения требования в источнике. Если говорить о Борином примере, то это период исправной работы локомотива. Обозначим среднюю продолжительность исправной работы $1/\beta$. Если на линии (в источнике) находится в какой-то период k локомотивов, то интенсивность потока, поступающего в систему, равна k/β. Обслуживание требований пусть производится по показательному закону, средняя продолжительность обслуживания $1/\mu$. Общее число требований, циркулирующих в системе, N. Давайте изобразим граф возможных состояний такой системы (рис. 2.19). Состояние x_k ($k=0,1, \ldots, N$) соответствует k требованиям в системе. Как вы, наверно, заметили, этот граф отличается от всех, которые мы рисовали раньше,

тем, что интенсивности переходов «слева направо», т. е. переходов, соответствующих приходу новых требований, неодинаковы и зависят от того, к какому состоянию они относятся. По известному нам мнемоническому правилу можно написать соответствующую этому графу систему уравнений:

$$\begin{cases} N\beta p_0 = \mu p_1, \\ [(N-k)\beta + \mu]\, p_k = (N-k+1)\beta p_{k-1} + \mu p_{k+1}; \quad 0<k<N; \\ \mu p_N = \beta p_{N-1}; \\ \sum\limits_{k=0}^{N} p_k = 1. \end{cases} \tag{2.64}$$

Число требований в очереди такой системы n, следовательно, число уравнений (2.64) ограничено, но ограничение это является не искусственным условием, а естественным следствием конечности числа требований в системе и источнике. Выразим все вероятности состояний через p_0. Для этого перепишем (2.64) так:

$$\begin{cases} p_1 = \dfrac{N\beta}{\mu} p_0; \\ p_2 = \dfrac{N(N-1)\beta^2}{\mu^2} p_0; \\ \cdots\cdots\cdots\cdots \\ p_k = \dfrac{N(N-1)\ldots(N-k+1)\beta^k}{\mu^k} p_0; \quad 0<k<N; \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ p_N = \dfrac{N(N-1)\ldots 1\beta^k}{\mu^N} p_0; \\ \sum\limits_{k=0}^{N} p_k = 1. \end{cases} \tag{2.65}$$

Если мы просуммируем все уравнения (2.65), кроме последнего, и прибавим p_0, то, воспользовавшись последним уравнением, получим

$$p_0 = [1 + N(\beta/\mu) + N(N-1)(\beta/\mu)^2 + \ldots + N(N-1)\ldots \\ \ldots 1(\beta/\mu)^N]^{-1}. \qquad (2.66)$$

Таким образом, уравнения (2.65) и (2.66) полностью выражают вероятности состояний замкнутой системы через параметры потока требований и обслуживания. Вы можете сравнить эти формулы с выражениями для аналогичной разомкнутой системы (2.37) и (2.38) и убедиться в том, что расчет для разомкнутой системы делать значительно легче.

— Вот теперь, пожалуй, мы поговорили с вами о всех основных типах систем массового обслуживания. Конечно, есть много других случаев, которые и на практике встречаются, и в теории массового обслуживания рассматриваются. Но они, как правило, носят частный характер, и с ними вы сможете познакомиться, если сами заглянете в специальную литературу. Вы хотите что-то спросить, Боря?

— Скажите, пожалуйста, а рассматриваются ли в рамках теории массового обслуживания какие-нибудь задачи оптимизации?

— Рассматриваются, и даже в нескольких аспектах. Я расскажу только о некоторых подходах к задачам оптимизации в массовом обслуживании. Даже при рассмотрении самой простой одноканальной системы может возникнуть вопрос, какие распределения длительности интервалов входящего потока и обслуживания при фиксированных средних значениях дают наименьшие простои требований в очереди. Из формулы Полячека—Хинчина (2.56) для системы с простейшим

входящим потоком и из верхней оценки (2.59) для произвольного потока видно, что чем меньше дисперсии входящего потока и обслуживания, т. е. чем меньше «случайности», тем меньше ожидают требования в системе. В предельном случае, когда и интервалы входящего потока и интервалы обслуживания детерминированы, очереди вообще никогда не бывает, и требования начинают обслуживаться сразу же после прибытия. Отсюда можно сделать вывод: даже если не в ваших силах изменить входящий поток, старайтесь сделать интервалы обслуживания как можно более близкими друг другу, и требования в вашей системе будут мало простаивать.

— Вот любопытно, — воскликнула Галя, — из нижней оценки (2.60) для средней длины очереди получается, что даже если нагрузка в системе очень мала, например 0,001, а дисперсия интервалов обслуживания огромна, то и очередь при такой ничтожной нагрузке будет огромной. Неужели это возможно?

— Конечно. При малой нагрузке очередь может быть сколь угодно велика. Но для этого входящий поток должен быть «уродом»: его интервалы имеют громадный разброс. На практике мы никогда с такими потоками не встречаемся, поэтому ваше открытие и показалось вам странным.

Но эта оптимизация, так сказать, элементарна. Она преследует цель просто сократить время ожидания. Более сложные задачи оптимизации возникают там, где имеется некоторый экономический функционал, включающий доходы от работы системы и затраты на поддержание ее нормального функционирования. В системах с несколькими потоками простои требований из разных потоков могут обходиться дороже или дешевле, то же самое может относиться и к стоимостям обслуживания разных требований. В оптимальном режиме

такой системы некоторые требования, быть может, целесообразно пропустить раньше, другие стоит задержать. Здесь возникают задачи об оптимальном назначении приоритетов, решение этих задач — одна из важных проблем специалистов, которые занимаются приоритетными системами массового обслуживания.

Такую оптимизацию можно назвать статической в том смысле, что она проводится перед созданием системы, при ее проектировании. Однако в массовом обслуживании рассматривается ряд задач, где оптимизация производится постоянно в процессе функционирования системы. Это задачи об оптимальном управлении системами массового обслуживания. Но прежде чем говорить об оптимальном управлении, нужно сказать о том, что такое управление системой обслуживания.

Помните, мы говорили о системе с резервными приборами. Эти приборы автоматически включались, как только возникала очередь. Теперь представьте, что включение и выключение прибора связаны с какими-то расходами. Вы можете управлять системой и включаете дополнительные приборы по своему усмотрению. Пусть образовалась небольшая очередь перед тем прибором, который работал с самого начала. Простой требований, ожидающих обслуживания, связан с расходами. В такой ситуации вы сравниваете, что вам выгоднее: продолжать обслуживание требований одним прибором и нести расход, связанный с простоем требований, или ускорить их прохождение через систему, включив дополнительный прибор, и при этом понести расход, связанный с его включением. Когда очередь достигнет некоторой величины, стоимость простоя превысит стоимость включения, и вы включите еще один прибор. Теперь предположим, что очередь уменьшается. Так как единица времени работы прибора имеет свою стоимость, при некоторой очереди стоимость простоя имеющихся в

ней требований может стать меньше стоимости эксплуатации двух приборов, и вы выключите второй прибор. Но если и выключение прибора имеет свою стоимость, вы можете задержать его, имея в виду возможность появления новых требований, из-за которых вновь придется проводить включение и нести дополнительные расходы. Этот пример — типичная управляемая система. Она имеет стоимости простоя требований, работы приборов, их включения и выключения. Управление вы стараетесь выбрать таким образом, чтобы некоторый функционал, включающий все эти стоимости, минимизировался (если речь идет о расходах) или максимизировался (если речь идет о доходах). Если вам удалось выбрать управление, при котором этот функционал достигает своего экстремума, можно говорить об оптимальном управлении.

В описанной системе происходит изменение скорости обслуживания требований путем подключения дополнительных приборов. Прибор может быть и один, но иметь «набор» различных скоростей обслуживания (причем более быстрое обслуживание обходится дороже). Принципиальной разницы здесь нет. Частным случаем является система с одним прибором, одной скоростью обслуживания и возможностью включения и выключения. Выключение ничего не стоит и производится, когда система оказывается пустой, а включение связано с расходами. Прибор имеет смысл включать не сразу после прихода первого требования, а спустя некоторое время, когда очередь достигнет определенной длины. В этом случае прибору обеспечена продолжительная работа без выключения. Такая организация работы приводит к сокращению числа включений и связанных с ними расходов. С другой стороны, очередь, которая собирается перед включением прибора, не должна быть слишком большой, иначе расходы, связанные с просто-

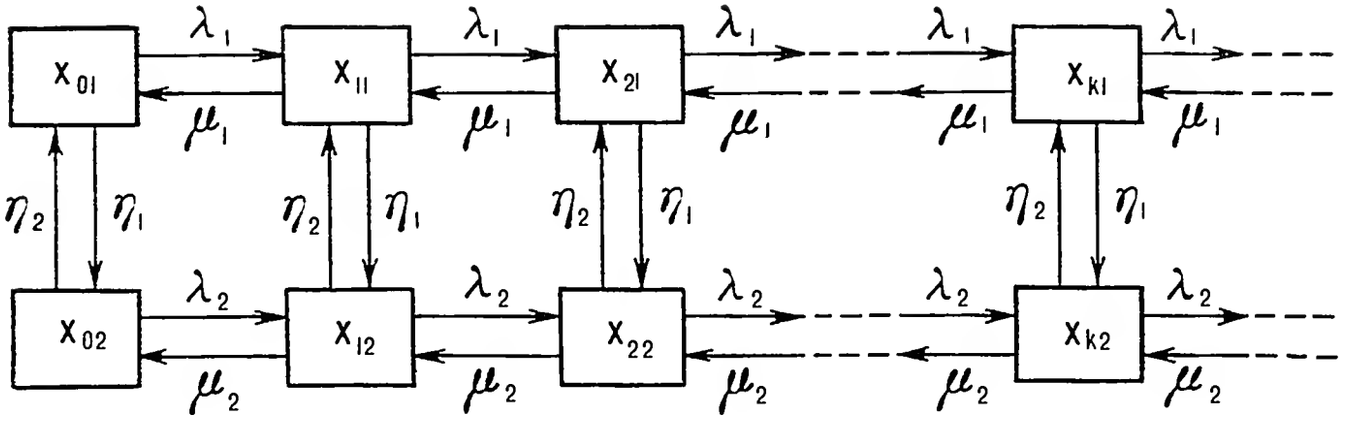


Рис. 2.20

ем требований, превзойдут экономию от сокращения числа выключений. Возникает задача об отыскании оптимальной длины очереди, при которой следует производить включение прибора.

Существует много других задач управления системами массового обслуживания — управляемыми могут быть порядок выбора из очереди на обслуживание, интенсивность входящего потока, распределение приоритетов между требованиями разных типов. В некоторых случаях в управлении принимают участие сами требования: они могут присоединяться к очереди или нет, если им это невыгодно.

Рассматриваются и так называемые системы с управляющим центром. Под центром понимается некоторая другая система, работе которой подчинена работа управляемой системы. При изменении состояний управляющего центра в управляемой системе может изменяться, например, интенсивность входящего потока или интенсивность обслуживания. Давайте в качестве примера рассмотрим управляемую одноканальную систему с простейшим входящим потоком, показательным распределением длительности обслуживания и неограниченной очередью. Эта система управляется каким-то центром, который может принимать два состояния, переход между этими состояниями происходит с интенсивностями

η_1 и η_2. Если управляющий центр находится в состоянии 1, в системе массового обслуживания интенсивности входящего потока и обслуживания λ_1 и μ_1, а если в состоянии 2, то соответственно λ_2 и μ_2. На рис. 2.20 показан граф состояний этой системы. Состояние x_{k1} означает, что в управляемой системе k требований, а управляющий центр находится в состоянии 1. Состояние x_{k2} означает, что при том же числе требований управляющий центр находится в состоянии 2. Если все распределения в системе являются показательными, то можно написать уравнения для ее состояний и получить нужные характеристики.

— Мне кажется, то, что я рассказал вам, — сказал Александр Андреевич, отложив карандаш, — достаточно, чтобы представить, насколько, с одной стороны, сложна наука об очередях, а с другой стороны, как много практических приложений в различных областях она охватывает.

— Помните, — обратился Борис к Стрелкину, — вы сказали утром, что все очереди одинаковы?

— Сдаюсь, сдаюсь, — поднял руки Стрелкин. — Теперь я совсем так не думаю. Интересно, а есть ли какая-нибудь практическая польза от этой теории? Может быть, вы знаете?

— Практическая польза есть от любой математической теории, а от теории массового обслуживания тем более, поскольку эта область прикладной математики возникла из нужд практики. Без теории массового обслуживания немыслимы расчет и конструирование современных телефонных станций и линий связи, вычислительных систем и систем городского хозяйства. На железнодорожном транспорте с помощью теории массового обслуживания рассчитывают перед строительством или модернизацией технические средства на сортировочных станциях, на водном транспорте определяют не-

обходимое количество причалов в портах, на автомобильном — размеры станций технического обслуживания. Методы теории массового обслуживания проникли в такие, казалось бы, далекие области, как создание сетей искусственных спутников и организация службы скорой помощи, социология и размещение пожарных частей в городе, вычислительная техника и организация туристических экскурсий. Недаром теория массового обслуживания, возникшая немногим более полувека назад, насчитывает в настоящее время несколько десятков книг и несколько тысяч статей в научных журналах. Чем сложнее будут технические и социальные системы, с которыми столкнется наука в своем развитии, тем больше будет нуждаться она в математических теориях, и не в последнюю очередь — в теории массового обслуживания.

БЕСЕДА ТРЕТЬЯ

Чем сложнее, тем проще

(Динамика средних)

Сельский пейзаж за окном сменился видами городской окраины. Поезд застучал на стрелках и из-под его колес побежали в сторону новые рельсы. Они сходились и расходились, не решаясь, как будто, отходить далеко от основного пути, а потом резко ушли в сторону и, растекаясь, заполнили открывшееся пространство. Всюду, насколько было видно, стояли вагоны, тысячи вагонов, между ними сновали казавшиеся маленькими тепловозы.

— Сортировочная станция, — кивнул на окно Стрелкин. — Между прочим, не самая большая.

— Да-а, сложное у вас хозяйство, — сказал Александр Андреевич, давно уже молча смотревший в окно. — Столько вагонов, и каждый нужно отправить по назначению. Впрочем, — он улыбнулся, — я слыхал, что вагоны иногда попадают и не по адресу?

— Случается, но редко. И каждый такой случай — ЧП. У нас есть тщательно разработанная система контроля и проверок, так что стараемся работать без брака. Сортировочная станция — это как большой почтамт: собираем вагоны, сортируем их и отправляем по разным адресам. Конечно, хозяйство у нас побольше и посложнее, чем на почте, и несведущему человеку покажется странным, как можно управиться со всей этой махиной, но железнодорожники делают это уверенно. И, между прочим, как я Вам уже говорил, иногда обходятся без математики.

— Как же вы это делаете?

— Практика, голая практика. Годами, десятилетиями вырабатывались правила, составлялись подробнейшие инструкции. Вот по ним и работаем. Ну, и, конечно, опыт.

— Но ведь с годами ваши станции увеличивались и усложнялись, железнодорожная сеть расширялась, движение становилось интенсивнее. Рано или поздно, в лю-

бой области практической деятельности наступает момент, когда старые методы становятся неэффективными. Предположим, что вас еще устраивают давно составленные инструкции, но не задумывались ли вы о том, что когда-нибудь, и, быть может довольно скоро, возникнет необходимость в новых, более современных способах управления?

— Вы, наверно, имеете в виду автоматизированное управление, АСУ? — спросил Борис. — Я знаю, что автоматизированные системы управления внедряются на железных дорогах, на некоторых сортировочных станциях уже стоят вычислительные машины. Но чем здесь помогает математика?

— Дело в том, что разработка сложной системы, особенно такой, как система управления, не может обойтись без предварительного математического расчета.

— Да, но математическому расчету поддаются довольно простые системы, если же система сложна, математика оказывается бессильной.

— Вот в этом я с вами согласиться не могу, — улыбнулся Александр Андреевич. — Знаете ли вы, что иногда чем сложнее система, тем проще и точнее ее можно рассчитать?

— Не может быть! Весь мой, может быть, небольшой опыт убеждает меня в противном.

— Слышали ли вы о методе динамики средних?

— Кажется, не приходилось. А ты, Галя?

— Нет.

— И я в первый раз слышу, — сказал Стрелкин. — Интересно, о чем это?

— Динамика средних — это сравнительно молодая область прикладной математики, которая занимается изучением сложных систем, состоящих из большого числа элементов. Давайте рассмотрим, например, локомо-

тивы, обслуживающие движение поездов на полигоне сети. Ежедневно на участки выходят десятки локомотивов. Для простоты будем считать все локомотивы одинаковыми, и наша система будет состоять из однородных элементов. Вообще же методы динамики средних применимы и для систем с неоднородными элементами. Но продолжим.

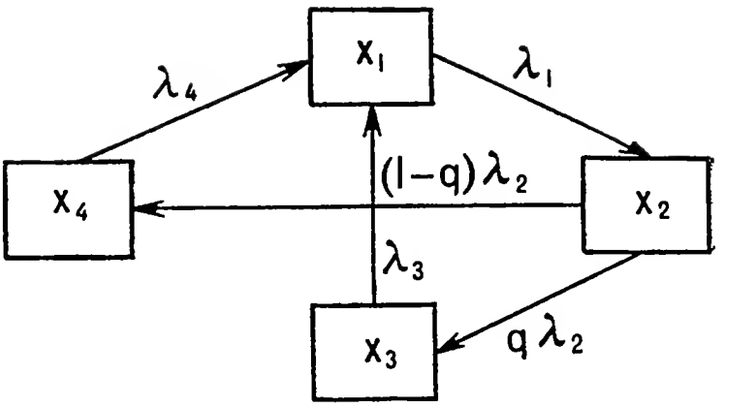


Рис. 3.1

— Каждый элемент системы — локомотив — может находиться в одном из нескольких состояний. Укажите нам их, пожалуйста, — обратился к Стрелкину Александр Андреевич.

— В каких состояниях может находиться локомотив? — задумчиво переспросил Стрелкин. — Их немного: исправное состояние — раз. Если случается поломка, возможны варианты: ремонт на месте и ремонт в депо.

— На определение неисправности уходит время — это тоже состояние?

— Да, поиск неисправности — второе состояние. В некоторых случаях после определения неисправности приступают к ремонту на месте.

— Скажем так: с вероятностью q производится ремонт на месте.

— Хорошо. С вероятностью q ремонт производится на месте. Это третье состояние. После окончания ремонта локомотив возвращается в первое состояние — продолжает работу на линии. С вероятностью $(1—q)$ ремонт на месте произвести нельзя. Локомотив следует в депо и там его ремонтируют — это четвертое состояние. После третьего и четвертого состояний локомотив возвращается в первое.

— Посмотрим на рисунок (рис. 3.1), — сказал Александр Андреевич. — Здесь я изобразил граф возможных состояний каждого элемента системы — локомотива. Давайте сразу проставим около стрелок интенсивности соответствующих переходов — так мы делали в задачах массового обслуживания.

— А что здесь понимать под интенсивностью перехода?

— Назовем интенсивностью выхода из состояния x_i величину, обратную среднему времени нахождения элемента в этом состоянии, — обозначим ее λ_i. Если из состояния x_i только один выход, то λ_i и будет интенсивностью перехода. Когда выходов несколько, то если по некоторому выходу производится переход с вероятностью q, интенсивность соответствующего перехода будет $q\lambda_i$. Предположим, что в нашем примере все λ_i и q известны — эти величины можно получить при достаточно длительном наблюдении за работой исследуемых объектов. Как нам найти характеристики одного элемента?

— Какие характеристики надо искать?

— Ну, например, предельные вероятности состояний p_1, p_2, p_3 и p_4 для стационарного режима.

— А как распределены продолжительности нахождения элементов в разных состояниях?

— Это, конечно, очень существенно. Давайте для простоты будем считать, что распределения всех случайных величин — показательные.

— Тогда мы имеем дело с марковским случайным процессом? — спросил Борис.

— Да, действительно, в системе с показательными распределениями времени нахождения во всех состояниях протекает марковский процесс и, следовательно, для составления уравнений относительно вероятностей состояний p_1, p_2, p_3 и p_4 можно использовать то самое мнемоническое правило, которым мы так часто пользова-

лись, когда говорили о массовом обслуживании. Давайте его вспомним: «Предельная вероятность i-го состояния, умноженная на сумму интенсивностей потоков, переводящих систему (в нашем случае это пока один элемент) в другие состояния, равна сумме произведений предельных вероятностей состояний, из которых можно перейти в i-е за один шаг, на интенсивности этих переходов». Итак, напишем для графа состояний элемента, изображенного на рис. 3.1, систему уравнений, не забывая при этом, что сумма вероятностей равна единице:

$$\begin{cases} \lambda_1 p_1 = \lambda_3 p_3 + \lambda_4 p_4; \\ \lambda_2 p_2 = \lambda_1 p_1; \\ \lambda_3 p_3 = q\lambda_2 p_2; \\ p_1 + p_2 + p_3 + p_4 = 1. \end{cases} \qquad (3.1)$$

— Вы забыли уравнение $\lambda_4 p_4 = (1-q)\lambda_2 p_2$, — подсказала Галя.

— Это уравнение нам не нужно — у нас уже есть четыре уравнения для четырех неизвестных. Оно линейно зависит от остальных уравнений, — сказал Борис. — Мы уже сталкивались с таким случаем.

— Я не совсем понимаю, о чем идет речь, — вмешался Стрелкин. — Александр Андреевич собирался рассказать о новом методе исследования сложных систем со многими элементами, а вместо этого мы говорим о вещах, нам уже хорошо известных: как решить задачу для одного элемента.

— Немного терпения, — улыбнулся Александр Андреевич. — По новому методу уравнения для всей системы получаются из уравнений для одного элемента. Какие характеристики системы будут нас интересовать?

— По-видимому, надо знать, сколько локомотивов работают, сколько ремонтируют в депо и на линии,

сколько стоит неисправными, — подумав, ответил Стрелкин.

— Иными словами, нас будут интересовать численности состояний X_1, X_2, X_3 и X_4; X_i означает число элементов, находящихся в состоянии x_i. При этом

$$X_1 + X_2 + X_3 + X_4 = N, \tag{3.2}$$

где N — число всех элементов, в нашем случае — локомотивов. Численности состояний X_i — случайные величины. Нас будут интересовать математические ожидания $m_i = M[X_i]$ и дисперсии $D_i = D[X_i]$ численностей состояний.

Совершенно ясно, что если проводить исследование обычными методами, то надо решать систему, состоящую из очень большого числа уравнений. Действительно, мы знаем, что для каждого состояния пишется свое уравнение, и количество состояний огромно: это количество возможных комбинаций состояний всех элементов. Мы будем действовать по-другому. Введем вспомогательные величины X_i^m, которые определим так:

$$X_i^m = \begin{cases} 1, & \text{если } m\text{-й элемент находится в состоянии } x_i \ (m=1, 2, \ldots, N; \\ & \quad i = 1, 2, 3, 4). \\ 0, & \text{если не находится.} \end{cases}$$

Ясно, что $\sum_{m=1}^{N} X_i^m = X_i$

и

$$m_i = M[X_i] = M\left[\sum_{m=1}^{N} X_i^m\right] = \sum_{m=1}^{N} M[X_i^m]. \tag{3.3}$$

Так как отдельные элементы независимы между собой,

$$D_i = D[X_i[= D\left[\sum_{m=1}^{N} X_i^m\right] = \sum_{m=1}^{N} D[X_i^m]. \tag{3.4}$$

Каждый элемент в произвольный момент времени находится в состоянии x_i с вероятностью p_i (и, следовательно, не находится с вероятностью $1 - p_i$). Следовательно,

$$M\left[X_i^m\right] = 0(1 - p_i) + 1p_i = p_i \tag{3.5}$$

и

$$D\left[X_i^m\right] = (0 - M\left[X_i^m\right])^2(1 - p_i) + \\ + (1 - M\left[X_i^m\right])^2 p_i = p_i(1 - p_i). \tag{3.6}$$

Так как элементы однородны, из (3.3) и (3.4) следует:

$$m_i = Np_i; \tag{3.7}$$

$$D_i = Np_i(1 - p_i). \tag{3.8}$$

Вернемся к системе уравнений (3.1) для одного элемента. Умножим все уравнения на N

$$\begin{cases} \lambda_1 Np_1 = \lambda_3 Np_3 + \lambda_4 Np_4; \\ \lambda_2 Np_2 = \lambda_1 Np_1; \\ \lambda_3 Np_3 = q\lambda_2 Np_2; \\ Np_1 + Np_2 + Np_3 + Np_4 = N. \end{cases} \tag{3.9}$$

— Теперь все ясно, — сказал Борис. — Если подставить уравнение (3.7) в (3.9), получится система уравнений для средних численностей состояний:

$$\begin{cases} \lambda_1 m_1 = \lambda_3 m_3 + \lambda_4 m_4; \\ \lambda_2 m_2 = \lambda_1 m_1; \\ \lambda_3 m_3 = q\lambda_2 m_2; \\ m_1 + m_2 + m_3 + m_4 = N. \end{cases} \tag{3.10}$$

Решить такую систему не составляет труда. А вот как составить систему уравнений для дисперсий D_i?

— Дисперсии D_i определяются совсем просто, система уравнений даже не требуется. Достаточно подставить уравнение (3.7) в (3.8):

$$D_i = m_i\left(1 - \frac{m_i}{N}\right). \qquad (3.11)$$

— Теперь, — подвел итог Александр Андреевич, — принцип применения метода динамики средних для системы, состоящей из однородных элементов с показательными распределениями, вам ясен: по обычным правилам составляется система уравнений для вероятностей состояний одного элемента, только вместо самих вероятностей стоят математические ожидания численностей этих состояний.

— То, что вы рассказали, Александр Андреевич, просто, — сказала Галя. — Но, наверно, в динамике средних есть и более сложные задачи?

— Конечно, есть. То, что я рассказал, — очень простой пример. Давайте его немного усложним и одновременно приблизим к реальности. Предположим, что некоторые локомотивы время от времени списываются. Естественно считать, что это может произойти, когда элемент находится в состоянии x_2. Будем теперь говорить, что неисправный локомотив с вероятностью q_1 ремонтируется на линии, с вероятностью q_2 ремонтируется в депо и с вероятностью $(1-q_1-q_2)$ списывается. Одновременно происходит процесс пополнения численности состояний — прибывают новые локомотивы и попадают в состояние x_1. Интенсивность этого пополнения обозначим γ. Будем считать, что в среднем число прибывающих новых элементов равно числу списываемых элементов. Таким образом, величина γ, будучи при-

веденной к одному элементу, равна $(1-q_1-q_2)\lambda_2 p_2$. На рис. 3.2 показан граф этой системы. Напишем уравнения сразу для средних численностей состояний

$$\begin{cases} \lambda_1 m_1 = \lambda_3 m_3 + \lambda_4 m_4 + (1 - q_1 - q_2)\lambda_2 m_2; \\ \lambda_2 m_2 = \lambda_1 m_1; \\ \lambda_3 m_3 = q_1 \lambda_2 m_2; \\ \lambda_4 m_4 = q_2 \lambda_2 m_2. \end{cases} \tag{3.12}$$

Обратите внимание на то, что здесь мы не воспользовались правилом $\sum_{i=1}^{4} m_i = N$, так как в некоторые моменты времени оно может не выполняться.

— Александр Андреевич, — сказал Борис, — вы говорили, что метод динамики средних позволяет рассчитывать систему с большим числом элементов, причем чем больше элементов, тем точнее расчет. Между тем, из того, что вы рассказали, не видно никакой зависимости между числом элементов и точностью результатов. Этот метод одинаково хорошо применим и для двух и для тысячи элементов. В чем дело?

— В разобранном нами примере действительно нет зависимости точности расчета от числа элементов. Здесь мы получили точный результат. Дело в том, что, во-первых, мы предположили, что все распределения случайных величин — показательные, а, главное, мы считали, что работа одного элемента не зависит от других. На практике же такие «идеальные» системы встречаются очень редко. Как правило, предельные вероятности состояний одного

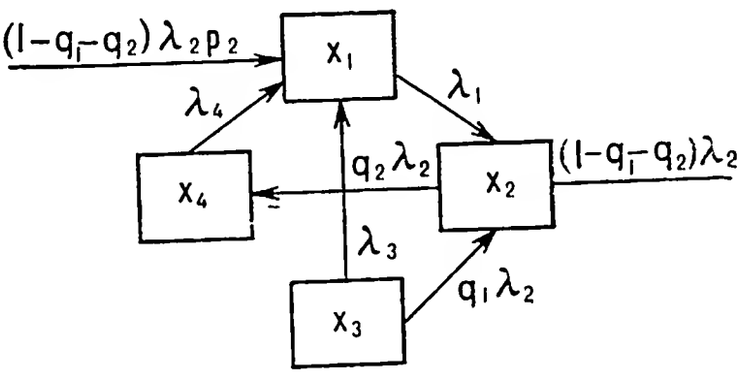


Рис. 3.2

элемента зависят не только от его собственных характеристик, но и от работы всей системы.

Вернемся к нашему примеру с локомотивами. Обратите внимание на состояние x_4. Мы считали, что среднее время ремонта в депо $^1/\lambda_4$. При этом мы негласно полагали, что ремонтные бригады готовы немедленно приступить к исправлению неисправности. На самом же деле это так только при условии, что в состоянии x_4 находится мало других локомотивов и имеются свободные ремонтные бригады. Если, например, в депо всего лишь одна ремонтная бригада, то для выполнения этого условия требуется, чтобы к моменту поломки локомотива ни один другой локомотив не находился в состоянии x_4. Если нет свободных ремонтных бригад, локомотив ожидает в очереди, время этого ожидания входит во время нахождения в состоянии x_4 и влияет на интенсивность λ_4. Эта «добавка» зависит от численности состояния x_4, причем характер этой зависимости нам неизвестен.

— Получается, что такая задача неразрешима! — воскликнула Галя.

— Точное решение получить действительно не удается, но можно найти приближенное решение, — ответил Александр Андреевич. — Для этого нужно сделать допущение, которое называется «принципом квазирегулярности»; считать, что интенсивности потоков событий, переводящих элемент из состояния в состояние, зависят не от самих численностей состояний, а от их математических ожиданий m_i. Поясним действие этого принципа на примере. Для простоты изложения возьмем упрощенную схему работы элемента — локомотива. Будем считать, что он может находиться в одном из двух состояний: x_1 — исправен, работает на линии, x_2 — неисправен, ремонтируется. Если бы ремонтных бригад было столько же, сколько локомотивов, все они ремонтировались бы независимо друг от друга, и граф состояний

элемента имел бы вид, показанный на рис. 3.3, *а*. В этом случае λ_1 и λ_2 были бы постоянными величинами и не зависели от численностей состояний. Система уравнений динамики средних, соответствующая рис. 3.3, *а*, при условии показательности распределений времени нахождения элемента в x_1 и x_2 имеет вид:

$$\begin{cases} \lambda_1 m_1 = \lambda_2 m_2; \\ m_1 + m_2 = N. \end{cases} \qquad (3.13)$$

Теперь предположим, что имеются три ремонтные бригады. Интенсивность переходов из неисправного состояния в исправное уже не является постоянной величиной, а зависит от числа ремонтируемых элементов. Обозначим эту переменную интенсивность $\tilde{\lambda}_2$ (граф состояний элемента — рис. 3.3, *б*).

Для того чтобы определить интенсивность $\tilde{\lambda}_2$, приходящуюся на один элемент, определим вначале зависимость сум-

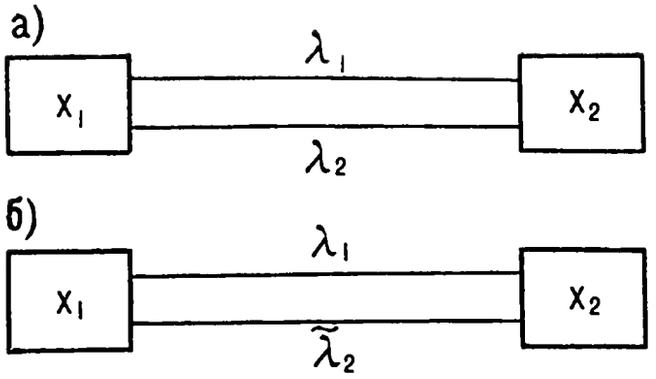


Рис. 3.3

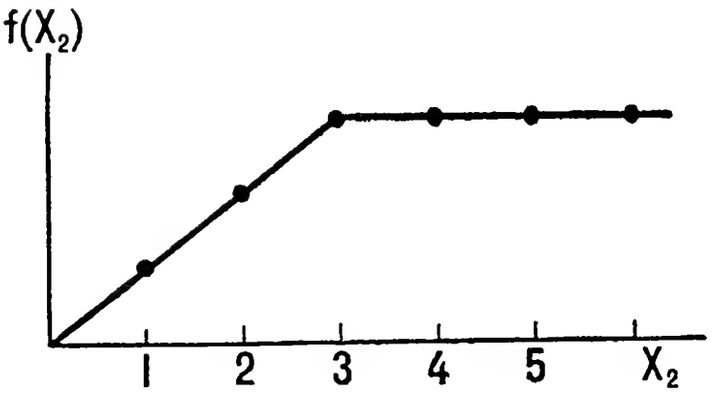


Рис. 3.4

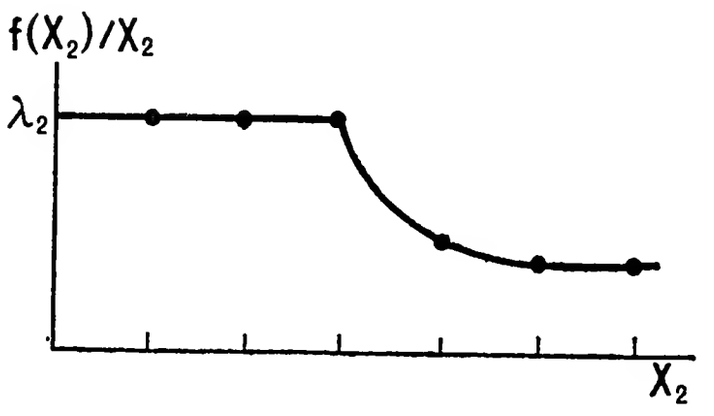


Рис. 3.5

марной интенсивности переходов элементов из x_2 в x_1 от X_2 — численности состояния x_2. Эта зависимость $f(X_2)$ показана на рис. 3.4. При $X_2=3$ интенсивность ремонтов достигает наибольшего значения — работают все три бригады. При дальнейшем увеличении X_2 интенсивность ремонта возрасти уже не может. График на рис. 3.4 для удобства изображен в виде непрерывной зависимости, хотя на самом деле эта зависимость дискретная.

Теперь подсчитаем, какая интенсивность ремонта приходится на один элемент. Для этого суммарную интенсивность $f(X_2)$ надо разделить на X_2 — численность состояния ремонта.

$$\widetilde{\lambda}_2 = f(X_2)/X_2. \tag{3.14}$$

На рис. 3.5 показан вид зависимости $\widetilde{\lambda_2}$ от X_2. Применяя принцип квазирегулярности, т. е. считая, что интенсивность перевода элемента из x_2 в x_1 зависит не от численностей состояний, а от их средних значений, сделаем замену X_2 на m_2 — математическое ожидание численности состояния ремонта

$$\widetilde{\lambda}_2^0 = f(m_2)/m_2. \tag{3.15}$$

Теперь уравнения динамики средних имеют вид:

$$\begin{cases} \lambda_1 m_1 = f(m_2); \\ m_1 + m_2 = 1, \end{cases} \tag{3.16}$$

где зависимость $f(m_2)$ задана графиком на рис. 3.5.

— Какова же точность этого приближенного метода? — спросил Борис.

— Точность зависит от числа элементов N и вида зависимости $f(X_2)$. Чем ближе функция f к линейной, тем больше точность. Что касается общего числа эле-

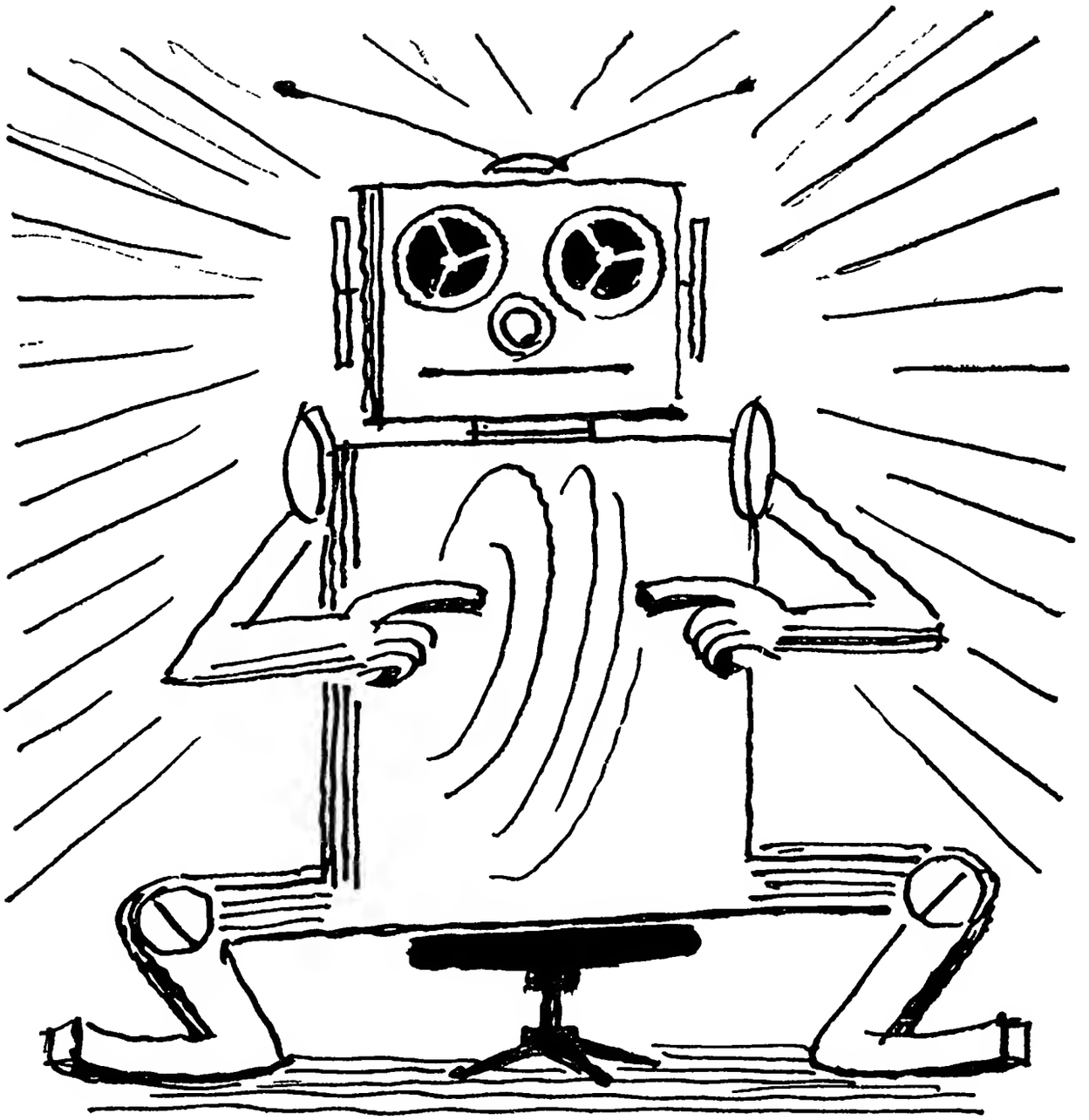
ментов, то точность возрастает с увеличением их числа N. В разных задачах эта зависимость различна. Но на основании опыта можно сказать, что погрешность будет очень мала, если число элементов равно нескольким сотням; если число элементов — несколько десятков, погрешность, как правило, тоже будет находиться в допустимых пределах. Иногда, когда функции близки к линейным, достаточно, чтобы $N=10$.

— Но это при условии, что все распределения случайных величин показательные, — напомнил Борис.

— Не обязательно. Дело в том, что, как показывает практика использования метода динамики средних, он хорошо работает и в случае произвольных распределений. Здесь я, к сожалению, ничего не могу доказать, а ссылаюсь лишь на практику. Раз распределения случайных величин произвольны, мы формально не можем записать уравнение динамики средних, однако большое число элементов, массовость явления делают вид распределений случайных величин не очень существенным. Поэтому при практическом использовании метода динамики средних часто даже не проводят статистического исследования работы элементов для выяснения характера распределений, а сразу пишут уравнения так, как будто все распределения показательные.

— Что ж, действительно оказывается, что чем сложнее система, чем больше в ней элементов, тем точнее она рассчитывается, — сказал Борис.

— И еще на одну особенность этого метода я хочу обратить ваше внимание, — добавил Александр Андреевич. — Строгими математическими приемами пока не удалось доказать правомерность его использования в широком классе задач. Пока что это сделано с помощью метода Монте-Карло, о котором мы вскоре поговорим.



БЕСЕДА ЧЕТВЕРТАЯ

Бросим жребий

(Метод Монте-Карло)

— Все-таки аналитические методы применимы для весьма ограниченного класса систем, — сказал Борис. — Система должна быть довольно простой, а случайный процесс, протекающий в ней, как правило, должен быть марковским. Конечно, для многих сложных систем можно построить упрощенные модели, поддающиеся аналитическому исследованию, но при этом мы часто не знаем величину погрешности, которая возникает при такой замене. И здесь можно «с водой выплеснуть ребенка», т. е. отказаться от каких-то существенных связей и зависимостей, которые, возможно, оказывают решающее влияние на работу всей системы.

— Такой расчет может принести только вред, — заметил Стрелкин.

— В сложных случаях нас выручит метод статистических испытаний, или, как его еще называют, метод Монте-Карло, — сказал Александр Андреевич. — Этот метод позволяет получить характеристики работы систем практически любой сложности. С помощью метода Монте-Карло можно оценить величину погрешности, возникающей при использовании формул, которые либо получены аналитически (с предварительным упрощением модели), либо предложены эмпирически. Идея метода очень проста: вы моделируете, т. е. повторяете работу реальной системы в искусственных условиях и исследуете модель так, как вы исследовали бы работу настоящего объекта: собираете статистику, обрабатываете ее, делаете соответствующие выводы, даете рекомендации.

— Зачем же строить модель? — спросила Галя. — Можно исследовать сам объект.

— Самого объекта может еще не быть, — вы как раз должны его проектировать. Но, предположим, объект существует, вам нужно изменить параметры его работы с целью улучшения, — например, достроить до-

полнительные пути на сортировочной станции, причем это можно сделать в разных вариантах, и надо выбрать лучший. Естественно, вы не будете пробовать все варианты по очереди. Вот тут и нужна модель, на ней можно опробовать все возможные изменения. Кроме того, благодаря применению ЭВМ для статистического моделирования скорость процесса, протекающего в модели, может в тысячи раз превосходить скорость процесса, протекающего в моделируемой системе. Процесс, происходящий в течение многих лет, может быть «сжат» в несколько минут.

В основе метода Монте-Карло лежат несколько ключевых понятий: алгоритм, единичный жребий, реализация. Алгоритм задает все возможные изменения, происходящие в системе. Например, при моделировании работы сортировочной станции алгоритм учитывает возможности прихода новых составов, появления в них неисправных вагонов, выхода из строя локомотивов, роспуска составов и т. д. Поскольку процесс в реальной системе случайный, все случайности должны быть воспроизведены и в модели. Делается это с помощью единичного жребия: каждый раз, когда в системе должно с некоторой вероятностью произойти одно или несколько событий, производится некоторое действие, аналогичное бросанию жребия, и в зависимости от исхода этого «бросания» выбирается соответствующее действие. Последовательность таких решений составляет реализацию работы модели. Так как каждая реализация складывается из цепи случайных событий, разные реализации отличаются друг от друга, но каждая воссоздает реальную ситуацию, которая могла бы встретиться на настоящем объекте.

Ключевым моментом в методе Монте-Карло является получение случайного числа, равномерно распределенного на отрезке [0, 1]. В настоящее время имеют-

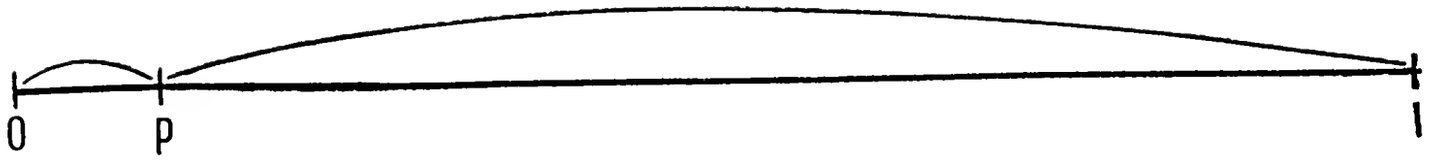


Рис. 4.1

ся десятки так называемых датчиков случайных чисел, задачей которых является выработка на ЭВМ чисел, равномерно распределенных на отрезке [0, 1]. Датчик случайных чисел представляет собой небольшую программу, которая позволяет получать распределение, в большей или меньшей степени приближающееся к равномерному. Как правило, точность такого приближения прямо пропорциональна сложности программы.

— Как используются эти случайные числа? — спросила Галя.

— Вот несколько примеров. Во время технического осмотра каждый вагон может оказаться неисправным с вероятностью p и исправным с вероятностью $1-p$. Как алгоритм определит, считать ли вагон исправным или нет? Он обратится к датчику случайных чисел. Если полученное число меньше p, алгоритм будет считать вагон неисправным, если больше — исправным (рис. 4.1).

Другой пример. Вагон спускается с горки. Он может попасть на один из 20 путей сортировочного парка. С вероятностью $p_i (i=1, 2, ..., 20)$ вагон попадает на i-ый путь. Как алгоритм выберет, на какой путь отправить очередной вагон? Он обратится к датчику; если число r, выданное датчиком, $0<r<p_1$, вагон «отправляется» на первый путь, если $p_1<r<p_1+p_2$ — на второй, если $p_1+p_2<r<p_1+p_2+p_3$ — на третий и так далее (рис. 4.2).

В этих примерах решались довольно простые вопросы — типа «быть или не быть» или «один из многих».

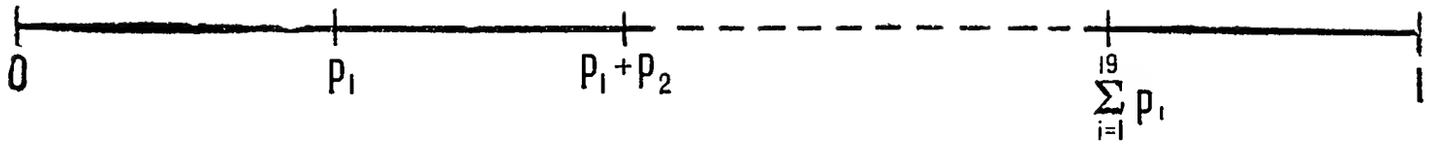


Рис. 4.2

Очень важным в методе Монте-Карло является получение непрерывно распределенных случайных величин. В том же примере с сортировочной станцией нужно получать значения интервалов между моментами прихода поездов, интервалов их роспуска и так далее. Причем законы распределения этих случайных величин могут быть самыми разными, и надо уметь получать их, используя датчик случайных чисел, т. е. равномерное распределение на отрезке [0, 1]. Здесь нам на помощь приходит одна теорема, суть которой состоит в следующем.

Пусть имеется функция распределения $F(x)$ непрерывной случайной величины. На рис. 4.3 эта случайная величина положительна, вообще же это не обязательно. $F(x)$, как известно, принимает значения на отрезке [0, 1]. Возьмем функцию, обратную $F(x)$, обозначим ее $F^{-1}(r)$. Оказывается, что если брать числа r, равномерно распределенные на отрезке [0, 1], то значения $F^{-1}(r)$ являются случайными числами, имеющими функцию распределения $F(x)$. На рис. 4.3 графически показана процедура получения чисел с заданным законом распределения.

— Нельзя ли пояснить это на конкретном примере? — попросил Стрелкин.

— Давайте посмотрим, как вырабатываются случайные величины с показательным распределением:

$$F(x)=\begin{cases} 1-e^{-\lambda x} & \text{при } x \geqslant 0; \\ 0 & \text{при } x < 0. \end{cases} \qquad (4.1)$$

Возьмем обратную функцию F^{-1}. Тогда величина

$$X = -\frac{1}{\lambda}\ln(1-r), \quad (4.2)$$

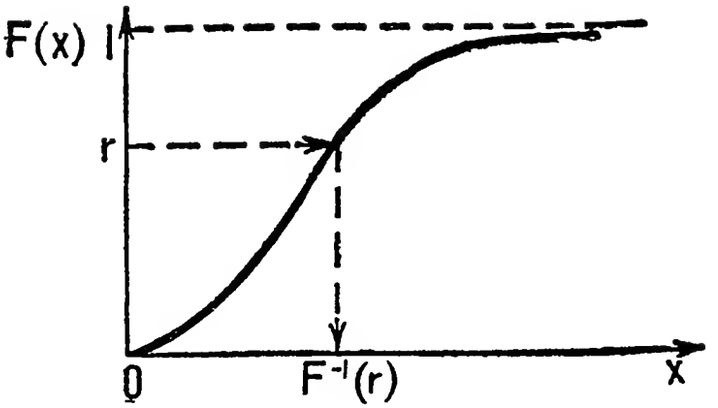


Рис. 4.3

где r — случайное число, равномерно распределенное на отрезке [0, 1]. Величина X распределена по показательному закону с параметром λ. Так как $(1-r)$ и r имеют одинаковое распределение, то выражение для получения показательно распределенных случайных величин можно записать в виде

$$X = -\frac{1}{\lambda}\ln r. \quad (4.3)$$

— Следовательно, для каждого закона распределения, который мы хотим моделировать, нужно вычислить обратную функцию для функции распределения,— сказал Борис. — Но, насколько я понимаю, в общем случае это не такая легкая задача.

— Действительно, часто нелегко найти обратную функцию. Но безвыходных ситуаций не бывает. Я сейчас приведу несколько примеров, показывающих, как можно без этого обойтись.

Плотность распределения случайной величины, распределенной по закону Эрланга k-го порядка, имеет уже знакомый нам вид:

$$f_k(x) = \begin{cases} \dfrac{\lambda(\lambda x)^{k-1}}{(k-1)!}e^{-\lambda x} & \text{при } x \geqslant 0; \\ 0 & \text{при } x < 0. \end{cases} \quad (4.4)$$

Для закона Эрланга нелегко написать функцию распределения, найти же еще обратную функцию чрезвычайно сложно. Поэтому действуют другим способом. Воспользуемся тем, что случайная величина, распределенная по закону Эрланга k-го порядка, может быть представлена как сумма k независимых случайных величин, каждая из которых распределена по показательному закону с параметром λ. Об этом мы упоминали в разговоре о теории массового обслуживания. Следовательно, случайные величины, распределенные по закону Эрланга, можно получать по формуле

$$X = -\frac{1}{\lambda}\sum_{i=1}^{k} \ln r_i. \tag{4.5}$$

Для каждого X нужно получить k равномерно распределенных случайных чисел.

Гиперэкспоненциальное распределение. Плотность такого распределения имеет вид

$$f(x) = \begin{cases} \sum_{i=1}^{k} p_i \lambda_i e^{-\lambda_i x} & \text{при } x \geqslant 0 \left(\sum_{i=1}^{k} p_i = 1 \right); \\ 0 & \text{при } x < 0. \end{cases} \tag{4.6}$$

Мы остановимся на случае $k=2$

$$f(x) = \begin{cases} p\lambda_1 e^{-\lambda_1 x} + (1-p)\lambda_2 e^{-\lambda_2 x} & \text{при } x \geqslant 0; \\ 0 & \text{при } x < 0. \end{cases} \tag{4.7}$$

Случайную величину, распределенную по гиперэкспоненциальному закону второго порядка, можно представить в таком виде: с вероятностью p выбирается

случайная величина, имеющая показательное распределение с параметром λ_1, а с вероятностью $(1-p)$ выбирается случайная величина, имеющая показательное распределение с параметром λ_2. Отсюда ясна схема моделирования.

Вырабатываются две величины, равномерно распределенные по [0, 1]: r_1 и r_2. Если
$0<r_1<p$,

то

$$X = -\frac{1}{\lambda_1}\ln r_1; \tag{4.8}$$

если

$$p < r_1 < 1,$$

то

$$X = -\frac{1}{\lambda_2}\ln r_2. \tag{4.9}$$

Плотность нормального распределения имеет вид

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}}, \tag{4.10}$$

где a — математическое ожидание;
σ — среднее квадратическое отклонение.

Если взять нормированную сумму n одинаково распредельных случайных величин ω_i с математическим ожиданием, равным m, и дисперсией D

$$\gamma = \frac{\sum_{i=1}^{n} \omega_i - nm}{\sqrt{nD}}, \tag{4.11}$$

то по центральной предельной теореме при $n\to\infty$ γ распределена по нормированному нормальному закону. Нормальное распределение с математическим ожида-

нием a и средним квадратическим отклонением σ можно получить, используя соотношение

$$\beta = \sigma\gamma + a. \tag{4.12}$$

Таким образом, величину

$$\beta = \frac{\left(\sum_{i=1}^{n} \omega_i - nm\right)\sigma}{\sqrt{nD}} + a \tag{4.13}$$

при достаточно больших n можно считать распределенной по нормальному закону с математическим ожиданием a и средним квадратическим отклонением σ. Вопрос, таким образом, сводится к тому, какое n считать «достаточно большим».

Практика показывает, что хорошее приближение к нормальному закону получается уже при $n=6$. Если в качестве ω_i взять случайные величины r_i с равномерным распределением на [0, 1] (в этом случае $m=\frac{1}{2}$, $D=\frac{1}{12}$, $\sqrt{nD}=\frac{1}{2}$), то нормально распределенные случайные величины можно получить так:

$$X = \sqrt{2}\left(\sum_{i=1}^{6} r_i - 3\right)\sigma + a. \tag{4.14}$$

— Пока что мы говорили о «кирпичиках», из которых складывается моделирование, — продолжал Александр Андреевич. — Самым ответственным моментом в методе статистических испытаний является разработка схемы моделирования, или алгоритма, по которому будет воссоздаваться работа моделируемой системы. Давайте рассмотрим весьма простой случай: моделиро-

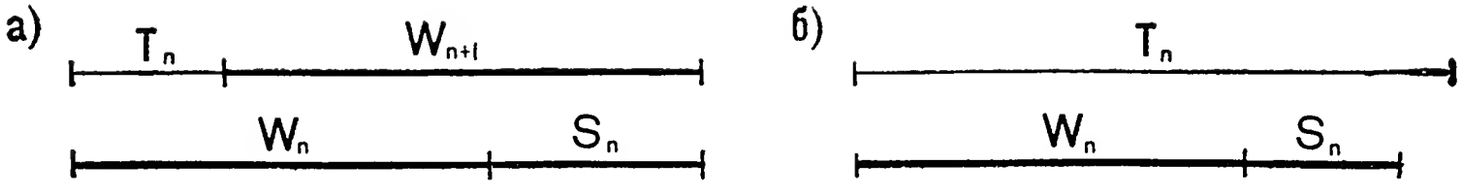


Рис. 4.4

вание одноканальной системы массового обслуживания. В разговоре о теории массового обслуживания мы упоминали о том, что если в систему поступает произвольный входящий поток и продолжительность обслуживания распределена по произвольному закону, точного аналитического решения для такого случая нет. С помощью же метода Монте-Карло сравнительно легко можно получить интересующие нас характеристики с достаточной степенью точности.

Обозначим T_n, S_n, W_n — соответственно интервал между моментами прихода n-го и $(n+1)$-го (с момента начала моделирования) требований, продолжительность обслуживания n-го требования и время его ожидания в очереди. Тогда, как видно из рис. 4.4, время ожидания $(n+1)$-го требования равно

$$W_{n+1} = W_n + S_n - T_n$$

(рис. 4.4, *а*), если интервал между поступлениями n-го и $(n+1)$-го требований меньше времени нахождения (W_n+S_n) n-го требования в системе. В противном случае $W_{n+1}=0$ (рис. 4.4, *б*). Таким образом, можно записать:

$$W_{n+1} = \begin{cases} W_n + S_n - T_n & \text{при } W_n + S_n > T_n;\ (n > 0); \\ 0 & \text{при } W_n + S_n \leqslant T_n. \end{cases} \qquad (4.15)$$

Рекуррентное соотношение (4.15) и является основой для составления схемы моделирования одноканальной

системы. Естественно, возникает вопрос: чему должно быть равно W_1? Эта величина выбирается по желанию того, кто производит моделирование. Например, можно считать $W_1=0$. Но если по каким-то соображениям известно приближенное значение среднего времени ожидания требований $W_{оч}$, лучше положить W_1 равным этому значению. Схема моделирования будет выглядеть так:

1. Задать W_1;
2. $n=1$;
3. Получить T_n;
4. Получить S_n;
5. Вычислить W_{n+1} по схеме (4.15);
6. Если n достигло заданного числа циклов моделирования, произвести обработку статистики и закончить.

Если нет, то

7. $n=n+1$;
8. Идти к п. 3.

В пунктах 3 и 4 величины T_n и S_n подсчитываются так, как мы говорили. Например, если распределения интервалов входящего потока и обслуживания показательные, T_n и S_n должны быть вычислены по формуле (4.5) (естественно, с разными λ для T и S).

Пункт 6 требует особого разъяснения. Во-первых, каким должно быть число циклов моделирования, или, другими словами, приход какого числа требований нужно задать? Легко заметить, что чем дольше моделировать, тем точнее будут результаты обработки статистики. С другой стороны, увеличение продолжительности моделирования приведет к большим затратам машинного времени. Это весьма существенно, так как нас, как правило, интересует, как будет работать система при различных условиях, и, чтобы ответить на этот вопрос, процесс моделирования нужно повторить много раз. Общей рекомендации дать нельзя: для моделиро-

вания одной системы может быть достаточно тысячи циклов моделирования, для другой понадобятся десятки тысяч. Можно предложить такой достаточно простой прием. Через определенное число циклов моделирования проводить требуемую обработку статистики. Например, в результате моделирования нам хотелось бы знать среднее время ожидания требований в очереди $W_{оч}$. Будем подсчитывать эту величину через 50 циклов моделирования, т. е. при $n=50, 100, 150, \ldots$. Делается это чрезвычайно просто: при получении очередного значения W_{n+1} в пункте 5 следует прибавить эту величину к содержимому специально выделенной ячейки $\sum_{i=0}^{N} W_{n+1}$, а затем вычислять $\sum_{i=0}^{49} W_{n+1}/50$, $\sum_{i=0}^{99} W_{n+1}/100$, $\sum_{i=0}^{149} W_{n+1}/150$ и так далее. Значение этих величин через каждые 50 циклов следует выводить на печать, а моделирование нужно прекратить тогда, когда эти значения практически перестанут изменяться, т. е. результат «установится». Существуют и более тонкие методы определения количества циклов моделирования, или, как еще говорят, «длины реализации», но они довольно сложны, и для их изложения требуется дополнительный математический аппарат.

Конечно, в разговоре мы затронули лишь самые элементарные вопросы, связанные со статистическим моделированием, и рассмотрели простейший пример. При моделировании сложных систем, в которых одновременно могут происходить разные процессы, возникают проблемы, преодоление которых связано со значительными трудностями. Развитие метода Монте-Карло неразрывно связано с распространением вычислительной техники.

БЕСЕДА ПЯТАЯ

Какие решения принимать?

(Математическое программирование и теория игр)

Борис обратился к Александру Андреевичу:

— Вот мы с вами уже очень долго говорим о математике. Поняли, как можно математически описывать разные процессы, узнали, что такое моделирование на вычислительных машинах. Но пока я еще не совсем представляю себе, как же использовать всю эту математику для того, чтобы принять оптимальное решение. Когда мы говорили о моделировании, вы, правда, сказали, что, изменяя различные параметры модели, мы можем выбрать такие их значения, которые влекут за собой лучшие результаты. Но ведь такой перебор значений параметров ограничен и временем и возможностями ЭВМ. Где же уверенность, что мы выберем наилучшее решение из всех возможных, а не только из тех, что мы перепробовали?

— Да, это, конечно, самый важный вопрос при обсуждении математических моделей — может ли данная модель помочь нам в выборе наилучшего из всех возможных решений? Как действовать в той или иной ситуации? Это, так сказать, «вопрос всех вопросов». Как только человек задал себе этот вопрос — он уже не просто участник или наблюдатель процесса, он — «исследователь операции». И раздел математики, изучающий этот вопрос, так и называется «исследование операций». Раньше да зачастую еще и сейчас, этот раздел отождествляют с методами оптимизации. Но это не совсем одно и то же. Исследование операций включает в себя как составную часть методы оптимизации, но не ограничивается только ими.

— Как же так? — удивился Боря. — Ведь вы сами сказали, что исследование операций занимается проблемой выбора наилучшего решения, а теперь говорите, что это не то же самое, что выбор оптимального решения. Ведь методы оптимизации, судя по названию, для этого и существуют?

— Давайте немного подождем с ответом на этот вопрос. Может быть, через некоторое время, когда мы уже разберемся, что такое «оптимизация» и что такое «наилучшее решение», ваше недоумение, Боря, пройдет. А пока, обратимся к очень простой задаче. Пусть, например, из города A в город B нужно отправить некий груз, причем есть два варианта: можно везти его поездом или самолетом, что, конечно, будет дороже, но зато груз к получателю поступит значительно быстрее. Скажите мне, пожалуйста, как вы будете отправлять груз?

— Ну конечно же поездом, — рассмеялся Стрелкин, — ведь это дешевле.

— А если это скоропортящийся груз? Или очень срочно требующийся в городе B? — возмутилась Галя. — Вон сколько мы уже едем поездом, а самолетом мы с Борей были бы давно на месте.

— Ну-ну, не сердитесь, Галя, — улыбнулся Александр Андреевич, — наш патриот железных дорог все прекрасно понимает и сказал это лишь в шутку. Но главное, мы видим, что понятие оптимального решения связано с точкой зрения на то, что есть оптимальность, т. е. какой критерий оптимальности мы выберем. Если в нашем примере важна скорость, то мы, конечно, выберем самолет, если же нам нужно уменьшить затраты, — железную дорогу. Итак, об оптимальности можно говорить только с точки зрения какого-нибудь критерия. Иногда так и говорят — оптимизировать критерий.

— Постойте-ка, — вдруг нахмурился Стрелкин, — но ведь мы часто слышим, как говорят, например, «максимум продукции при минимуме затрат». Здесь же явно два критерия — объем продукции и величина затрат. Первый надо максимизировать, второй — минимизировать. А увеличение выпуска продукции всегда связано с дополнительными затратами. Как же так?

— Вы привели типичный пример многокритериальной задачи. Решать такие задачи непросто. Да и само понятие «решения» такой задачи не всегда четко определено. Вот мы пока и будем различать однокритериальные задачи и методы оптимизации, созданные для их решения, и задачи многокритериальные, являющиеся более общими задачами. Исследование операций включает в себя как те, так и другие задачи.

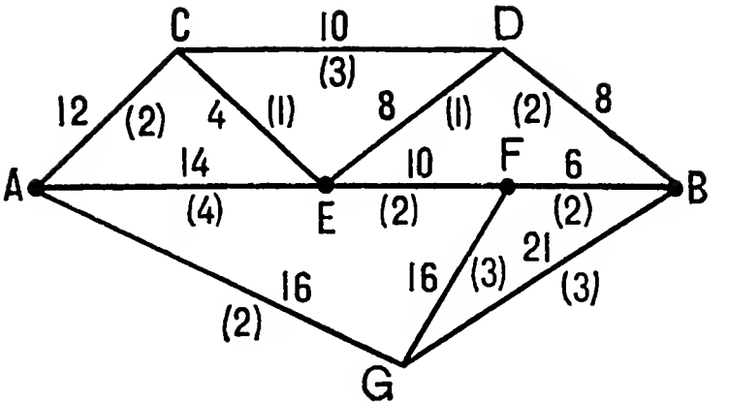


Рис. 5.1

Итак, обратимся к однокритериальным задачам. Возьмем такой пример. Опять из города A в город B нужно отправить груз, но между этими городами имеется сложная железнодорожная сеть, например, такая, как на рис. 5.1.

Первая цифра на отрезке, соединяющем два пункта, обозначает длину пути, а цифра в скобках — затраты на продвижение по этому пути.

Пусть сначала требуется найти кратчайший путь из A в B, т. е. нашим критерием является длина пути, по которому пойдет груз. Естественно, что кратчайший путь нужно искать только среди таких путей, которые не содержат петель, т. е. возвращений в уже пройденные пункты. Здесь таких путей 14. Вот они:

$$x_1 = ACDB;\ x_2 = ACEDB;\ x_3 = ACEFB;\ x_4 = ACEGFB;$$

$$x_5 = ACEGB;\ x_6 = AEDB;\ x_7 = AEFB;\ x_8 = AEGFB;$$

$$x_9 = AEGB;\ x_{10} = AGECDB;\ x_{11} = AGEDB;$$

$$x_{12} = AGEFB;\ x_{13} = AGFB;\ x_{14} = AGB.$$

— Да, но зачем рассматривать путь $ACEDB$, если путь $ACDB$ короче? — спросила Галя.

— Постойте, постойте, — ответил Александр Андреевич. — Мы пока еще только строим модель, а выбирать наилучшее решение будем позже. Для этого надо описать все наши действия. Пусть x обозначает какое-нибудь из возможных решений. Тогда каждому x отвечает $f(x)$ — длина пути x. Множество всех путей, выписанных выше, обозначим X. Теперь задача формулируется так: «минимизировать $f(x)$ на множестве X».

Нетрудно проверить, что в нашем случае наименьшее $f(x)$ равно 30 и достигается при $x_1=ACDB$, $x_7=AEFB$ или $x_6=AEDB$. И в качестве решения нашей задачи можно взять любое из этих трех значений x-путей.

— Да, но ведь затраты на путь $ACDB$ и на путь $AEDB$ меньше, чем на путь $AEFB$. Значит этот последний путь не годится, — возразил Стрелкин.

— Что значит — не годится? Мы ведь не накладываем никаких условий на стоимость проезда, — ответил Александр Андреевич. — Поэтому в той задаче, которую мы решали, все три пути равноправны и являются оптимальными с точки зрения критерия $f(x)$.

Если же мы теперь поставим задачу: среди всех кратчайших путей найти тот, который требует наименьших затрат, то тогда, конечно, путь $AEFB$ придется отбросить и оставить только пути $ACDB$ и $AEDB$.

Пусть теперь имеется ограничение: затраты не должны превышать 6. Обозначим затраты на путь x через $g(x)$. Тогда математически можно будет сформулировать нашу задачу так: «минимизировать $f(x)$ на множестве X при условии $g(x)\leqslant 6$».

Из рис. 5.1 видно, что только два пути удовлетворяют ограничению: $x_2=ACEDB$ и $x_{14}=AGB$, причем $g(x_2)=6$, а $g(x_{14})=5$. Решением задачи будет, очевидно, путь $x_2=ACEDB$, так как $f(x_2)=32$ и $f(x_{14})=37$.

На этом примере мы увидели, что однокритериальные задачи можно разбить на два класса — задачи безусловной оптимизации и задачи условной оптимизации.

— А как же задача с выбором самого дешевого пути из всех кратчайших? — спросил Борис.

— В той задаче фактически было два критерия, — быстро сказал Стрелкин, — ведь так?

— Вообще говоря, да, — ответил Александр Андреевич, — в этой задаче мы учитывали два критерия $f(x)$ и $g(x)$, которые надо было минимизировать. Но мы их рассматривали не одновременно, а последовательно, отдавая предпочтение критерию $f(x)$. Эту задачу тоже можно сформулировать как задачу условной оптимизации с одним критерием, например, так:

«минимизировать $g(x)$ на множестве X при условии

$$f(x)=\min_{x'\in X} f(x')»,$$

где $\min\limits_{x'\in X} f(x')$ означает наименьшее из всех $f(x')$ на множестве X.

Задачи условной оптимизации называют еще задачами математического программирования, причем условие может быть не одно. В общем виде эти задачи можно сформулировать так:

«минимизировать или максимизировать функцию $f(x)$ на множестве X при условии

$$g_1(x)\geqslant 0,\quad g_2(x)\geqslant 0,\ \ldots,\ g_m(x)\geqslant 0,$$

где $g_1(x)$, ..., $g_m(x)$ — некоторые функции, определенные на X».

Решать задачи условной оптимизации очень непросто. Для многих классов таких задач придуманы специальные методы нахождения оптимальных значений x.

Среди них самым первым был изучен класс задач линейного программирования. Это такие задачи условной оптимизации, где величины x описываются наборами определенного количества чисел и все входящие в их формулировку функции $f(x)$, $g_1(x)$,..., $g_m(x)$ линейны, т. е. $x=(x_1, \ldots, x_n)$ и

$$\begin{aligned} f(x) &= c_1 x_1 + \ldots + c_n x_n + c_0; \\ g_1(x) &= a_{11} x_1 + \ldots + a_{1n} x_n + a_{10}; \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ g_m(x) &= a_{m1} x_1 + \ldots + a_{mn} x_n + a_{in0}. \end{aligned} \tag{5.1}$$

К задачам линейного программирования сводятся многие практические задачи. Например, такая.

Пусть нам требуется организовать железнодорожные перевозки в некотором промышленном районе, в котором имеются, скажем, три пункта, производящих одинаковую продукцию, и четыре — потребляющих ее (например, три угольные шахты и четыре металлургических предприятия). Пусть в i-м пункте-производителе продукция выпускается в объеме A_i; $i=1, 2, 3$, а в j-м пункте-потребителе эта продукция требуется в объеме B_j, $j=1, 2, 3, 4$. Будем предполагать, что суммарное производство может удовлетворить весь спрос, т. е.

$$A_1 + A_2 + A_3 \geqslant B_1 + B_2 + B_3 + B_4.$$

Пусть транспортная сеть в районе такова, что каждый пункт-производитель связан с каждым пунктом-потребителем, но не все участки дороги требуют одинаковых транспортных затрат, что связано с длиной пути, с трудностью трассы, родом тяги и т. п.

Если затраты на перевозку единицы продукции из i-го пункта-производителя в j-й пункт-потребитель со-

ставляют c_{ij}, $i=1, 2, 3$, $j=1, 2, 3, 4$, и мы организуем перевозки так, что из i-го пункта в j-й будет доставляться x_{ij} продукции, то общие затраты составят

$$Z = c_{1,1}x_{1,1} + c_{2,1}x_{2,1} + c_{3,1}x_{3,1} + c_{1,2}x_{1,2} + \ldots + c_{3,4}x_{3,4}.$$

Естественно, что распределение перевозимой продукции, т. е. выбор x_{ij}, следует организовать так, чтобы величина Z была по возможности меньше. Таким образом, нам нужно минимизировать Z, выбирая x_{ij}. Но при этом x_{ij} не могут браться произвольно.

Во-первых, все величины x_{ij} должны быть неотрицательные. Во-вторых, из первого пункта-производителя не может быть вывезено больше, чем там производится, т. е.

$$x_{1,1} + x_{1,2} + x_{1,3} + x_{1,4} \leqslant A_1.$$

Аналогично:

$$x_{2,1} + x_{2,2} + x_{2,3} + x_{2,4} \leqslant A_2;$$
$$x_{3,1} + x_{3,2} + x_{3,3} + x_{3,4} \leqslant A_3.$$

И, наконец, чтобы удовлетворить спрос, необходимо доставить в первый пункт-потребитель не меньше чем B_1 продукции, т. е.

$$x_{1,1} + x_{2,1} + x_{3,1} \geqslant B_1$$

и аналогично:

$$x_{1,2} + x_{2,2} + x_{3,2} \geqslant B_2;$$
$$x_{1,3} + x_{2,3} + x_{3,3} \geqslant B_3;$$
$$x_{1,4} + x_{2,4} + x_{3,4} \geqslant B_4.$$

У нас получилась задача на поиск минимума функции Z при ограничениях. В таком виде эта задача не

похожа на общий вид (5.1) задачи линейного программирования, но ее можно легко переписать в нужной форме, если сделать замену неизвестных:

$$x_1 = x_{1,1};\ x_2 = x_{2,1};\ x_3 = x_{3,1};\ x_4 = x_{1,2}, \ldots,\ x_{12} = x_{3,4}.$$

Тогда, если положить $c_1 = c_{1,1}$, $c_2 = c_{2,1}$, ..., $c_{12} = c_{3,4}$, то мы получим

$$Z = c_1 x_1 + c_2 x_2 + \ldots c_{12} x_{12}$$

и ограничения примут вид:

$$x_k \geqslant 0,\ k = 1, \ldots,\ 12$$
$$x_1 + x_4 + x_7 + x_{10} - A_1 \leqslant 0;$$
$$x_2 + x_5 + x_8 + x_{11} - A_2 \leqslant 0;$$
$$x_3 + x_6 + x_9 + x_{12} - A_3 \leqslant 0;$$
$$x_1 + x_2 + x_3 - B_1 \geqslant 0;$$
$$x_4 + x_5 + x_6 - B_2 \geqslant 0;$$
$$x_7 + x_8 + x_9 - B_3 \geqslant 0;$$
$$x_{10} + x_{11} + x_{12} - B_4 \geqslant 0,$$

а это уже совсем просто приводится к виду (5.1), т. е. наша задача сформулирована как задача линейного программирования.

— А не может ли случиться так, — спросил Стрелкин, — что не будет ни одного x, которое удовлетворяет всем ограничениям?

— Конечно, может, — ответил математик. — В этом случае говорят, что задача математического программирования не имеет решения из-за того, что множество «допустимых» решений x, т. е. таких, которые

удовлетворяют всем ограничениям, пусто. Если это случилось при решении практической задачи, нужно изменить ограничения.

— Что значит — изменить ограничения? — спросила Галя. — Ведь они должны выполняться по условию, иначе бы мы их не писали!

— Дело в том, что ограничения, которые встречаются в практических задачах, часто носят не такой категорический характер: «нечто должно быть меньше данного значения». Обычно это выражается в том, что невыполнение такого ограничения влечет за собой большие потери, которые можно назвать «штрафом за нарушение ограничений». Этот штраф может выражаться различными способами.

Рассмотрим, например, такую задачу:

«Максимизировать $f(x)$ на множестве X при условии $g(x) \geqslant 0$».

Во-первых, введем такую функцию:

$$\varphi(x) = |g(x)| - g(x).$$

Очевидно, что $g(x) \geqslant 0$, тогда и только тогда, когда $\varphi(x) = 0$, так как при $g(x) \geqslant 0 \; |g(x)| = g(x)$ и наоборот. Кроме того, из неравенства $|a| \geqslant a$ следует $\varphi(x) \geqslant 0$ при всех x.

Теперь нашу задачу можно переформулировать в виде:

«Максимизировать $f(x)$ на множестве X при условии $\varphi(x) = 0$, где $\varphi(x)$ — функция, принимающая неотрицательные значения».

Будем теперь считать, что нам разрешается нарушить ограничение $g(x) \geqslant 0$, или, что то же самое, $\varphi(x) = 0$, но при этом придется платить штраф, тем больший, чем больше отличается $\varphi(x)$ от 0, т. е. если $\varphi(x)$ оказалось равным u, то мы должны заплатить штраф $L(u)$. Функция $L(u)$ называется штрафной функцией, она возрастает

с ростом u и $L(0)=0$. Самый простой вид штрафной функции линейный: $L(u)=\lambda u$, где $\lambda>0$. Теперь наша задача может быть переписана в виде:

«максимизировать $F(x)=f(x)-L(\varphi(x))$ на множестве X».

Здесь мы, конечно, подразумеваем, что $f(x)$ выражена в тех же единицах, что и «штраф» L.

Если x таково, что ограничение выполнено, т. е. $\varphi(x)=0$, то $F(x)=f(x)-L(0)=f(x)$.

Если же $\varphi(x)>0$, то $F(x)=f(x)-L(\varphi(x))<f(x)$.

Может быть, нарушив ограничение $\varphi(x)=0$, мы и получим большее $f(x)$, но зато заплатим штраф $L(\varphi(x))$ и «в сумме» получим не $f(x)$, а $F(x)=f(x)-L(\varphi(x))$.

— Что вы хотите сказать, Боря?

— Если я правильно понял, мы получили задачу оптимизации без ограничений.

— Совершенно верно, — подтвердил Александр Андреевич.

— Но, если у нас несколько ограничений $g_k(x)\geqslant 0$, $k=1,\ldots,m$, мы можем взять несколько функций $\varphi_k(x)$, неотрицательных и таких, что $g_k(x)\geqslant 0$, тогда и только тогда, когда $\varphi_k(x)=0$. Затем возьмем несколько штрафных функций $L_k(u)$ — да хотя бы $L_k(u)=\lambda_k u$ и задачу максимизации $f_k(x)$ при условии $g_k(x)\geqslant 0$, $k=1,\ldots,m$ заменим на задачу безусловной максимизации функции

$$F(x)=f(x)-\lambda_1\varphi_1(x)-\ldots-\lambda_m\varphi_m(x). \qquad (5.2)$$

А ведь найти максимум функции без всяких ограничений, наверно, легче, чем при ограничениях...

— Давайте по порядку, — остановил Борю Александр Андреевич. — Действительно, несколько ограничений можно так же просто «снять», введя штрафные функции, как и одно. Но при этом функция $F(x)$ оказывается много сложнее, чем функция $f(x)$, и не всегда безусловная оптимизация легче условной. Кроме того,

мы получили совсем новую задачу, допустив, что нарушать ограничения все же можно, хотя и с уплатой штрафа. В какой-то ситуации может оказаться, что выгоднее нарушить ограничение и заплатить штраф, но зато получить большее значение $F(x)$, чем, не нарушая ограничений, получить маленькое $f(x)$. Поэтому задача максимизации функции $F(x)$ и задача условной максимизации $f(x)$, вообще говоря, не одно и то же, и их решения могут не совпадать.

— Но неужели нельзя свести условную оптимизацию к безусловной, если функция (5.2) не очень сложная? — спросил Борис.

— В некоторых случаях можно. Предположим для простоты, что функция $f(x)$ при всех $x \in X$ ограничена одним и тем же числом M. Возьмем опять одно ограничение $g(x) \geqslant 0$ или $\varphi(x)=0$ и линейный штраф

$$F_\lambda(x) = f(x) - \lambda\varphi(x), \quad \lambda > 0.$$

Множество тех $x \in X$, для которых $\varphi(x)=0$, обозначим через D^*. Пусть z — максимальное значение $f(x)$ на множестве D^* и x^* — одно из тех в этом множестве, на которых максимум достигается, т. е. $f(x^*)=z$. Можно считать, что $M > z$.

Выберем $\lambda > 0$ и рассмотрим множество D_λ, состоящее из таких x, для которых $\varphi(x) \leqslant \frac{M-z}{\lambda}$. Если $x \notin D_\lambda$, то $\varphi(x) > \frac{M-z}{\lambda}$ и

$$F_\lambda(x) < f(x) - \lambda\,\frac{M-z}{\lambda} \leqslant M - (M-z) = z.$$

А так как D_λ включает в себя D^*, то, в частности, в D_λ есть точка x^*, в которой $F_\lambda(x^*)=f(x^*)=z$, ведь $\varphi(x^*)=0$. Значит при $\lambda > 0$ максимум функции $F_\lambda(x)$ не может

достигаться вне множества D_λ, и даже если мы никаких ограничений на x накладывать не будем, а будем искать безусловный максимум $F_\lambda(x)$ на множестве X, то решение все равно окажется в множестве D_λ.

Если $\lambda < \lambda'$, то $\frac{M-z}{\lambda} > \frac{M-z}{\lambda'}$ и те x, для которых $\varphi(x) \leqslant \frac{M-z}{\lambda'}$, будут удовлетворять и неравенству $\varphi(x) \leqslant \frac{M-z}{\lambda}$. Это означает, что с увеличением λ множество D_λ сужается ($D_{\lambda'} \subset D_\lambda$). Увеличивая λ до бесконечности, мы получим сужающуюся последовательность множеств D_λ, каждое из которых содержит D^*, и максимальное значение функции $F_\lambda(x)$ достигается только на x_λ, принадлежащих множеству D_λ. Если пересечение этих множеств не содержит никаких других, кроме принадлежащих D^*, и если функция $f(x)$ непрерывна, то, рассматривая точки x_λ, можно определить и точки x^*, на которых достигается максимум $f(x)$ в области D^*. Для вас, Боря и Галя, я могу сказать, что каждая предельная точка для последовательности x_λ будет служить решением задачи на условный максимум. Я хочу только добавить, что ограниченность функции $f(x)$ на всем множестве X, которую я сначала предположил и потом использовал в своих рассуждениях, несущественна. От этого предположения можно отказаться, выбирая специальным образом штрафные функции.

— Да, но ведь этак приходится решать бесконечно много задач безусловной оптимизации при разных λ, — разочарованно вздохнула Галя. — Уж лучше одна задача условной оптимизации.

— Во-первых, это не всегда так, — ответил Александр Андреевич. — Для некоторых задач можно подобрать одно такое достаточно большое λ, что всякое решение задачи безусловной оптимизации с использова-

нием штрафных функций является решением задачи условной оптимизации. Таковы, например, задачи линейного программирования. Но даже в тех случаях, когда приходится рассматривать целую последовательность задач безусловной оптимизации, это зачастую бывает оправдано. Поэтому метод штрафных функций стал сейчас очень быстро развиваться. В то же время, даже если удается свести задачу условной оптимизации к одной задаче безусловной оптимизации, это не всегда следует делать. Для тех же задач линейного программирования в настоящее время есть много методов, приводящих к оптимальному решению быстрее, чем методы решения задач безусловной оптимизации.

Мы не будем сейчас подробно останавливаться на различных задачах математического программирования и методах их решения. Посмотрим еще раз, что мы делали.

Сначала мы описали множество наших действий X и ограничения, которым они должны подчиняться. Затем выбрали критерий $f(x)$, имея при этом в виду, что, как только x выбрано, все дальнейшее определяется однозначно.

Тем самым мы построили математическую модель той реальной ситуации, которую изучаем. Дальше нахождение оптимального поведения в этой ситуации свелось к задаче математического программирования.

— Да, но так ведь почти никогда не бывает, — возразил Стрелкин.

— Что не бывает? — спросили одновременно Борис и Галя.

— Не бывает, чтобы мы могли точно предвидеть все последствия наших действий. В том же примере с посылкой грузов из A в B, где мы минимизировали длину пути и затраты, нельзя всегда точно сказать, каковы окажутся затраты на перевозку груза по данному

отрезку пути. Вдруг, например, окажется, что в связи с неправильным планированием эксплуатационной работы или в результате стихийных явлений (ливни, снегопады и т. д.) будет затруднено движение на данной линии и придется тянуть состав более мощным локомотивом. А в случае, если при этом груз из-за чего-то подпортится в пути, железная дорога еще и штраф будет платить. Да мало ли что может произойти!

— Верно, — ответил Александр Андреевич. — До сих пор мы считали, что условия — ограничения и критерий определены однозначно и можно точно сказать, что будет, если мы поступим тем или иным способом. На самом же деле все гораздо сложнее. И критерий $f(x)$ и функции $g_k(x)$ могут быть подвержены различным факторам, которые будем называть неконтролируемыми. Некоторые из этих факторов носят случайный характер, т. е. их можно рассматривать как случайные величины с известным законом распределения, а про некоторые мы вообще можем ничего не знать, кроме того, что они как-то влияют на результаты наших действий. Возьмем такой пример. Пусть мы хотим добраться из нашего старого доброго города A в не менее дорогой нам город B и у нас есть два пути — через город C и через город D. До городов C и D идут поезда. Из города C в B можно добраться только по озеру на теплоходе, а из D в B можно проехать поездом. Известно, что теплоход отправляется из C в B регулярно, каждый час, правда, мы не знаем точно, в какие моменты, но его скорость зависит от погоды (от ветра, например). А поезд из D в B отправляется то ли каждые 2 ч, то ли каждые 4 ч. Точно мы не знаем и спросить не у кого. Каким путем мы доберемся скорее?

Давайте сначала построим математическую модель ситуации. Ну-ка, Боря, какое здесь множество действий и какие факторы влияют на их результаты?

— Возможных действий здесь мало — выбрать путь через C или через D. Пусть это будут x^C и x^D. Погода — случайный фактор, можно определить, с какой вероятностью она будет благоприятна.

— Это, пожалуй, так, — согласился Александр Андреевич. — Пусть скорость теплохода будет $v_т+\xi$, где ξ — случайный фактор.

— Теперь, — продолжил Боря, — если мы выберем путь x^C, время путешествия будет зависеть от величины τ^C — интервала от прибытия поезда до отплытия теплохода, а если путь x^D, — то от величины τ^D — интервала от прибытия поезда в D из A до отправления поезда из D в B...

— Что же дальше? — спросил математик, — почему вы остановились?

— Потому что про τ^C мы можем кое-что сказать: $0 \leqslant \ll \tau^C \ll 1$, — ответил Боря. — А вот как быть с τ^D? Ведь может оказаться, что $0 \ll \tau^D \ll 2$ и $0 \ll \tau^D \ll 4$.

— Как так, — вмешался Стрелкин, — мы же не знаем, какой именно интервал между поездами из D в B — 2 или 4 ч. Поэтому мы должны считать, что $\tau^D \ll 4$. Это, наверно, и есть неконтролируемый фактор?

— Да, точно так же, как и τ^C, — сказал математик. — Давайте подведем итог. Будем считать, что поезда на всех участках идут с одинаковой скоростью $v_п$ и все расстояния AC, AD, CB и DB равны r. Наш критерий — длительность путешествия — будет зависеть от выбора x и от факторов ξ, τ^C и τ^D, связанных следующим соотношением, а именно:

$$f(x, \xi, \tau^C, \tau^D) = \begin{cases} \dfrac{r}{v_п} + \tau^C + \dfrac{r}{v_т + \xi}, & \text{если } x = x^C, \\ \dfrac{r}{v_п} + \tau^D + \dfrac{r}{v_п}, & \text{если } x = x^D. \end{cases}$$

Какой же путь выбрать, x^C или x^D?

— Можно, я попробую, — робко попросила Галя.

— Давайте, смелее, — подбодрил ее Александр Андреевич.

— Значит так, — начала Галя. — С ξ мы разделаемся просто. ξ — величина случайная, значит, и $f(x, \xi, \tau^C, \tau^D)$ — величина случайная. Поэтому мы вместо нее возьмем математическое ожидание

$$\overline{f}(x, \tau^C, \tau^D) = M\{f(x, \xi, \tau^C, \tau^D)\}.$$

Величина $\overline{f}$ будет показывать среднее время с учетом случайной погоды.

Теперь вместо τ^C и τ^D тоже возьмем их средние значения $\tau^C = \frac{1}{2}$ и $\tau^D = 2$. Получим $\overline{f}(x) = f\left(x, \frac{1}{2}, 2\right)$.

Осталось взять только то x, при котором $\overline{f}(x)$ меньше. Вот и все.

— Это почему же! — воскликнул Боря. — Почему мы должны брать средние значения τ^C и τ^D? Не лучше ли считать их самыми неблагоприятными и взять $\tau^C = 1$, а $\tau^D = 4$. Тогда мы гарантированно приедем в B не долее чем за $f(x, 1, 4)$ часа.

— А почему бы не взять тогда и ξ самым неблагоприятным? — спросил Стрелкин. — Тогда мы будем гарантированы от случайностей погоды.

— Да, как мы видим, здесь не все просто, — заметил Александр Андреевич. — Давайте разбираться по порядку. Итак, пусть у нас есть функция $f(x, \xi, \alpha)$, зависящая от нашего действия x, случайного фактора ξ и неконтролируемого фактора α, про который мы лишь знаем, что он принимает значения в множестве A. Для того чтобы перейти от критерия f к критерию $\overline{f}(x)$, не зависящему от ξ и α, мы должны поставить задачу бо-

лее четко, чем просто «минимизировать f». И здесь могут быть разные варианты. Чтобы все наши дальнейшие рассуждения сопровождались числовыми примерами, зададимся числами $r=100$ км, $v_п=60$ км/ч, $v_т=40$ км/ч, ξ может меняться от -10 км/ч до $+10$ км/ч, равновероятно принимая промежуточные значения на $[-10, +10]$, то есть ξ будет равномерно распределено на этом отрезке.

Теперь

$$f(x, \xi, \tau^C, \tau^D)=\begin{cases}\frac{5}{3}+\tau^C+\frac{100}{40+\xi}, & \text{если } x=x^C,\\ \frac{10}{3}+\tau^D, & \text{если } x=x^D.\end{cases}$$

Можно поставить такую задачу — какое минимальное время может занять у нас поездка при самых благоприятных условиях? В этом случае следует считать $\tau^C=\tau^D=0$ и $\xi=10$.

Тогда

$$\tilde{f}_1(x)=f(x, 10, 0, 0)=\begin{cases}3\frac{2}{3}, & \text{если } x=x^C,\\ 3\frac{1}{3}, & \text{если } x=x^D,\end{cases}$$

и минимальное значение равно $3\frac{1}{3}$, причем оно реализуется на пути x^D.

Ставя и решая такую задачу, мы можем только указать, что полученное значение f дает нам оценку снизу для всех возможных значений.

Математически эту величину можно записать так:

$$\min_x \tilde{f}_1(x)=\min_x \min_\xi \min_\alpha f(x, \xi, \alpha).$$

Станем теперь на другую крайнюю точку зрения, которую предложил нам Боря.

Будем считать все факторы неблагоприятными. Тогда:

$$\tau^C = 1,\ \tau^D = 4,\ \xi = -10$$

и

$$\tilde{f}_2(x) = f(x,\ -10,\ 1,\ 4) = \begin{cases} 6, \text{ если } x = x^C, \\ 7\frac{1}{3}, \text{ если } x = x^D. \end{cases}$$

В этом случае мы можем наверняка добраться из A в B за 6 ч, избрав путь через C. Эта величина, ее можно записать так: $\min\limits_x \tilde{f}_2(x) = \min\limits_x \max\limits_\xi \max\limits_\alpha f(x, \xi, \alpha)$, гарантирует нам путешествие не долее ее в любом случае.

Попробуем теперь сойти с этих крайних точек зрения, будем искать путь, который в среднем будет самым коротким. Для этого мы возьмем, как и предлагала Галя, математическое ожидание:

$$\tilde{f}(x, \tau^C, \tau^D) = M\{f(x, \xi, \tau^C, \tau^D)\} =$$

$$= \begin{cases} \frac{5}{3} + \tau^C + \int\limits_{-10}^{10} \frac{100}{40+\xi} \cdot \frac{d\xi}{20}, \text{ если } x = x^C, \\ \frac{10}{3} + \tau^D, \text{ если } x = x^D. \end{cases}$$

Здесь

$$\int\limits_{-10}^{10} \frac{100}{40+\xi} \cdot \frac{d\xi}{20} = 5 \ln \frac{5}{3} \approx 2{,}6.$$

Но вот как же быть с τ^C и τ^D?

Здесь можно опять стать либо на точку зрения «крайнего пессимизма» и считать $\tau^C = 1$, $\tau^D = 4$, либо постараться получить какую-нибудь информацию, уменьшающую неопределенность значений этих факто-

ров. Что же касается точек зрения «крайнего оптимизма» или «золотой середины», то они ничем не предпочтительнее любых других. «Крайний пессимизм» во всяком случае дает нам гарантированный результат. Но он может оказаться много хуже, чем результат, получаемый в действительности. Поэтому, если в математической модели участвуют какие-нибудь неконтролируемые факторы, лучше всего постараться о них получить побольше информации, сузив тем самым неопределенность их изменения и уменьшив их влияние на результат. Такую информацию можно получать, используя опрос сведущих специалистов-экспертов.

При отсутствии неконтролируемых факторов (когда есть лишь случайные факторы) задачи можно ставить, например, так:

1) максимизировать математическое ожидание $M\{f(x, \xi)\}$, выбирая x, удовлетворяющее условиям:

$$g_k(x) \geqslant 0, \quad k = 1, \ldots, m;$$

2) максимизировать вероятность $P\{f(x, \xi) \geqslant c\}$, где c — некоторая заданная величина при условиях:

$$g_k(x) \geqslant 0, \quad k = 1, \ldots, m;$$

3) максимизировать $M\{f(x, \xi)\}$, выбирая x так, чтобы вероятность выполнения ограничений $g_k(x, \xi) \geqslant 0$, $k = 1, \ldots, m$ была не меньше заданной величины P.

Таких постановок задач можно придумать множество. Они носят название задач стохастического программирования.

При этом можно в качестве решения задачи выбирать и x случайным.

— Как это — случайное x? — удивился Стрелкин. — Что, например, будет означать такое «решение» —

с вероятностью $\frac{1}{2}$ ехать налево, а с вероятностью $\frac{1}{2}$ — направо? Монетку, что ли, бросать?

— Конечно, сама по себе эта фраза звучит достаточно странно. Но представьте себе, что вы решаете одну и ту же задачу много раз и вас интересует суммарный результат. Тогда по закону больших чисел, если вы будете выбирать «обычные» неслучайные решения с определенными вероятностями, за достаточно большое число шагов N вы получите в сумме примерно NM, где M — математическое ожидание, отвечающее данному распределению вероятностей на X. Если вероятности были выбраны так, чтобы максимизировать M, то в результате вы и получите максимальную суммарную величину. Что же касается «монетки», то действительно можно воспользоваться «монеткой», чтобы реализовать последовательность неслучайных решений, приводящих к нужному результату. Только «монетка» эта должна быть такой, чтобы с нужной вероятностью «выпадало» нужное значение x. Такие «монетки» реализуются в ЭВМ с помощью датчиков случайных чисел. Мы уже упоминали их, когда говорили о моделировании.

Теперь обратимся к многокритериальным задачам. Здесь могут быть ситуации двух типов. Может оказаться так, что выбор действия сосредоточен в руках одного лица, которое руководствуется несколькими критериями и хочет их по возможности все максимизировать (если что-то нужно минимизировать, то можно рассматривать эту величину со знаком минус), т. е. он должен выбрать x из какого-то множества X некоторым «наилучшим» образом с точки зрения критериев $f_1(x),..., f_n(x)$. Но может оказаться так, что «действие» является коллективным в том смысле, что есть несколько лиц, каждый из которых выбирает свое действие x_i из некоторого множества X_i, $i=1, ..., n$, в резуль-

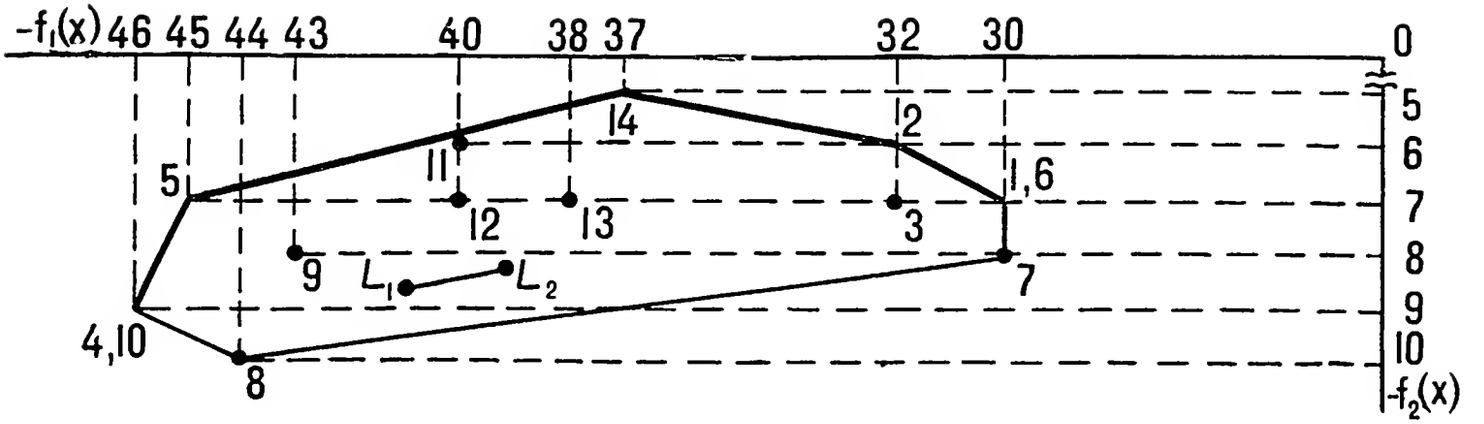


Рис. 5.2

тате чего этот i-й участник получает величину $f_i(x_1, ..., x_n)$, которую хочет максимизировать.

— Но разве это не одно и то же? — спросил Боря. — Ведь если мы обозначим набор $(x_1, ..., x_n)$ через x, то получится то же самое, что и в первом случае.

— Нет, не то же самое, — возразил Александр Андреевич. — Ведь для того чтобы выбрать какое-нибудь определенное $x=(x_1, ..., x_n)$, всем участвующим нужно договориться о совместных действиях, что, вообще говоря, не всегда возможно. Мы еще вернемся к этому. А пока рассмотрим первую ситуацию. Для простоты будем считать, что критериев всего два: $f_1(x)$ и $f_2(x)$. И знаете, что, давайте обратимся к нашему самому первому примеру о выборе пути. Там множество X состоит из 14 элементов. В качестве критериев возьмем длину пути и затраты, т. е. $f_1(x)=-f(x)$, $f_2(x)=-g(x)$. Нарисуем на плоскости оси координат, на которых будем откладывать значения $f_1(x)$ и $f_2(x)$ соответственно, когда x пробегает все множество X (рис. 5.2). У каждой точки поставим номер x, при котором получаются соответствующие значения $f_1(x)$ и $f_2(x)$.

Теперь ясно, что выбор нужно осуществить только из четырех путей x_1, x_2, x_6 и x_{14}, так как для любого другого пути найдется один, в котором оба критерия принимают не меньшее значение, а один строго большее.

Так, например,

$$(x_3)f < f_1(x_6); \quad f_2(x_3) = f_2(x_6);$$

$$f_1(x_7) = f_1(x_6); \quad f_2(x_7) < f_2(x_6);$$

$$f_1(x_{11}) < f_1(x_{14}); \quad f_2(x_{11}) < f_2(x_{14}).$$

Те решения x, для которых нельзя одновременно улучшить оба критерия, называются паретовскими по имени математика Парето. В нашем примере четыре паретовские точки и, не привлекая никаких дополнительных условий, мы не можем отдать предпочтение ни одной из них: решение x_{14} обеспечивает нам наименьшие затраты (5) при длине пути 37, а решение x_6 — кратчайший путь (30) при затратах, равных 7.

Теперь выбор «наилучшего» решения целиком находится в руках лица, принимающего решение, и носит неформальный характер. Модель только подсказывает, среди каких решений нужно делать этот выбор. Ситуация, как мы видим, отличается от задач простой оптимизации.

— А зачем вы соединили точки, отвечающие решениям x_{14}, x_2, x_6, x_7? — спросила Галя Александра Андреевича, внимательно рассматривая рис. 5.2.

— А вот зачем, — ответил тот. — Представим себе, что задачу о выборе пути мы решаем регулярно в течение длительного времени. Мы можем себе позволить каждый раз выбирать другой путь. Например, с вероятностью $^1/_3$ путь x_{14} и с вероятностью $^2/_3$ путь x_3. Тогда математическое ожидание первого критерия будет:

$$\frac{1}{3}f_1(x_{14}) + \frac{2}{3}f_1(x_3) = \frac{1}{3}(-37) + \frac{2}{3}(-32) = -33\frac{2}{3},$$

а второго

$$\frac{1}{3}f_2(x_{14})+\frac{2}{3}f_2(x_3)=\frac{1}{3}(-5)+\frac{2}{3}(-7)=-6\frac{1}{3}.$$

Точка с такими координатами делит отрезок, соединяющий точки, на рис. 5.2, отвечающие путям x_{14} и x_3 в отношении 1:2, считая от последней.

Вообще, если точка L_1 на рис. 5.2 соответствует пути x рисунка 5.1, а точка L_2 — пути x', то, взяв путь x с вероятностью p и x' с вероятностью $1-p$, мы получим значения математических ожиданий первого и второго критериев соответственно: $pf_1(x)+(1-p)f_1(x')$ и $pf_2(x)+(1-p)f_2(x')$.

Точка с такими координатами делит отрезок L_1L_2 в соотношении $p:(1-p)$, считая от точки L_2. Смысл таких «случайных» решений, которые в дальнейшем будем называть смешанными, в отличие от «обычных», которые будем называть простыми, или чистыми, мы уже выяснили. Если же мы будем «смешивать» не два, а три, четыре или больше «обычных» решений, то сможем получить весь многоугольник, с вершинами в соответствующих точках.

— Теперь понятно, — перебила его Галя. — Множество точек на рис. 5.2, оказавшееся внутри нарисованного многоугольника, есть всевозможные средние значения критериев f_1 и f_2 при использовании смешанных решений. Правильно?

— Да, но ведь при этом появились новые паретовские решения. Или я что-то не так понял? — спросил Борис.

— Нет, все верно, — ответил математик. — Действительно, у нас получился выпуклый многоугольник всех возможных исходов и вся его «северо-восточная» граница соответствует паретовским решениям — «смесям» чистых решений x_{14} и x_2 или x_2, x_1 и x_6.

Рис. 5.3

— Что же в этом хорошего?—спросил Стрелкин. — Был выбор из четырех путей, а стал — из бесконечно многих. Задача для выбора «наилучшего» решения усложнилась.

— Это не совсем так,— улыбнулся Александр Андреевич. — Ведь расширились наши возможности. Мы теперь можем выбирать различные «компромиссные» решения среди паретовских, которые раньше были невозможны. Не будем сейчас обсуждать, как именно можно выбирать ту единственную паретовскую точку, которую считают «наилучшей». Таких моделей выбора много. Эти модели называются моделями «торга». Они описывают как бы торговлю между критериями при выборе компромиссного решения, «справедливого» по отношению ко всем критериям.

Обратимся теперь к задачам со многими участниками. Этими задачами занимается раздел исследования операций, называемый теорией игр. Здесь и термины используются «игровые» — участники называются игроками, их действия — стратегиями, критерии — функциями выигрышей. Игроки могут образовывать коалиции и применять совместные согласованные стратегии. В частности, если все игроки объединятся в одну коалицию и будут действовать совместно, мы придем к задаче, которую только что рассматривали. Но здесь появляется кое-что новое. Пусть два игрока с множеством чистых стратегий X_1 и X_2 и функциями выигрышей $f_1(x_1, x_2)$ и $f_2(x_1, x_2)$ соответственно решили действовать совместно и договорились выбирать паретовское решение. На рис. 5.3 изображено множество всех возмож-

ных исходов в этой игре и выделено множество исходов, отвечающих паретовским решениям.

— А почему вы нарисовали это множество невыпуклым? — спросила Галя.

— Но ведь мы не говорим пока о смешанных стратегиях, — ответил Александр Андреевич. — Если бы игроки решили применять совместные смешанные стратегии, то множество исходов было бы выпуклым.

Итак, пусть игроки договорились о каком-то компромиссном паретовском решении (x_1^*, x_2^*). При этом первый получит выигрыш $f_1(x_1^*, x_2^*)$, а второй — $f_2(x_1^*, x_2^*)$.

Но теперь, подумав на досуге, какой-нибудь игрок, например первый, может захотеть нарушить договор и выбрать стратегию x_1, отличную от «договорной» x_1^*, если $f_1(x_1, x_2^*) > f_1(x_1^*, x_2^*)$.

Заметим, что при этом обязательно $f_2(x_1, x_2^*) < f_2(x_1^*, x_2^*)$, иначе решение (x_1^*, x_2^*) не было бы паретовским.

В задаче с одним лицом, принимающим решение, такой «обман» невозможен, так как принимающее решение лицо заинтересовано во всех критериях и уж если какое-то паретовское решение выбрано, оно будет реализовано. В игре каждый игрок заботится только о себе. Поэтому мы не можем упускать из виду возможность нарушений совместных соглашений. Для того чтобы совместное решение (x_1^0, x_2^0) было гарантировано от нарушений, достаточно, очевидно, выполнения неравенств:

$$f_1(x_1^0, x_2^0) \geqslant f_1(x_1, x_2^0) \quad \text{и} \quad f_2(x_1^0, x_2^0) \geqslant f_2(x_1^0, x_2)$$

при всех $x_1 \in X_1$ и $x_2 \in X_2$. Эти неравенства говорят о том, что никому из игроков невыгодно отклоняться от

решения (x_1^0, x_2^0). Такие решения называются равновесными решениями, или точками равновесия по Нэшу. Если среди паретовских решений есть равновесное, то, конечно, оно должно быть выбрано в качестве совместного решения. Беда в том, что часто среди паретовских решений нет равновесных. Да и не всегда равновесные решения существуют. Основная теорема теории игр, доказанная Нэшем, утверждает, что в случае, когда множества чистых стратегий всех игроков конечны и они применяют (независимо каждый свою) смешанные стратегии, то такая игра всегда имеет точку равновесия.

— А нельзя ли все это пояснить на простых примерах? — спросил Стрелкин.

— Конечно, сейчас мы так и поступим, — ответил Александр Андреевич.

Среди всех игр, которые можно описать, указав участников, их возможные действия и критерии — функции выигрышей, наиболее простым является класс

Таблица 5.1

i \ j	1 . . . . . j . . . . . n
1	a_{11} . . . . . a_{1j} . . . . . a_{1n}
.	. . . . .
.	. . . . .
.	. . . . .
i	a_{i1} . . . . . a_{ij} . . . . . a_{in}
.	. . . . .
.	. . . . .
.	. . . . .
m	a_{m1} . . . . . a_{mj} . . . . . a_{mn}

игр двух лиц с нулевой суммой, в которых каждый участник — игрок может принимать лишь конечный набор действий — чистых стратегий. Давайте подробнее разберем такие игры.

Будем обозначать через I множество всех чистых стратегий первого игрока, а J — второго. Оба эти множества конечны, пусть I состоит из m элементов (у первого игрока есть m возможных альтернатив), а J — из n элементов. Поскольку игра, которую мы рассматриваем, с нулевой суммой, т. е. сумма выигрышей игроков при любом выборе стратегий равна 0, то достаточно определить выигрыш первого игрока. Обозначим через a_{ij} выигрыш первого игрока в том случае, когда он применяет стратегию i, а его противник — j. Можно считать, что второй игрок платит первому сумму a_{ij} (если a_{ij} отрицательное, то это означает, что первый игрок платит второму). Такую игру удобно представить в виде следующей матрицы (табл. 5.1).

Поэтому игра и называется матричной.

Теперь мы можем попытаться выяснить, как следует вести себя игрокам в такой игре, т. е. определить их рациональные стратегии.

— Какие стратегии? — переспросил Стрелкин.

— Рациональные, те, которые разумно применять.

— А почему вы говорите так осторожно — рациональные? Почему не постараться найти оптимальные действия? Разве оптимальное поведение не есть рациональное?

— Все было бы очень просто, если бы мы могли всегда указать действия, реализующие тот или иной принцип оптимальности.

Вот в нашем случае, например, первый игрок стремится максимизировать свой выигрыш, а второй — минимизировать, так как он одновременно является проигрышем второго игрока. Здесь ситуация похожа на

перетягивание каната, и нет никакой гарантии, что существует какое-то оптимальное поведение у игроков, но ведь делать что-то надо. Какие же действия разумно применять? Вот какой смысл я придаю словам «рациональное» поведение, — ответил Александр Андреевич. — Сейчас все станет яснее. Давайте обратимся к примерам.

Пусть матричная игра задана табл. 5.2.

— Что можно посоветовать игрокам?

— Так сразу, наверно, не скажешь, — сказала Галя.

— Ну почему же, возразил Александр Андреевич.— Станем на точку зрения первого игрока. Как он расценит свои возможности?

— Ясно, что третью стратегию он никогда применять не будет, — сказал Стрелкин. — Ведь, что бы ни предпринял второй игрок, если первый выберет вторую стратегию, он всегда получит больше, чем если выберет третью. Верно?

— Совершенно верно. Это свойство называется доминированием. Если какая-то стратегия (здесь для первого игрока — третья) доминируется другой, то ее можно исключить из рассмотрения, так как в любом случае использовать доминируемую стратегию нерационально.

Таблица 5.2

i \ j	1	2	3
1	2	—1	—2
2	0	1	2
3	—1	0	1

Таблица 5.3

i \ j	1	2	3
1	2	—1	—2
2	0	1	2

— Значит, — заметил Борис, — нашу матрицу можно заменить меньшей, как на табл. 5.3, и ничего не изменится?

— Да, для первого игрока ничего не изменится. И в исходной игре и в той, матрицу которой сейчас написал Борис, рациональное поведение первого игрока будет одним и тем же.

— Ну хорошо, уберем мы все доминируемые стратегии, а что дальше? — спросил Стрелкин.

— А теперь будем рассуждать так. Если первый игрок выберет стратегию с номером i, то при выборе вторым игроком стратегии с номером j первый игрок получит a_{ij} и, следовательно, самое меньшее — $\underline{v}_i = \min\limits_{1 \leqslant j \leqslant n} a_{ij}$. Значит, выбор i-й стратегии гарантирует первому игроку выигрыш не меньший $\underline{v}_i$, и если первый игрок выберет стратегию с таким номером i_0, что

$$\underline{v}_{i_0} = \max_{1 \leqslant i \leqslant m} \underline{v}_i = \max_{1 \leqslant i \leqslant m} \min_{1 \leqslant j \leqslant n} a_{ij} = \underline{v},$$

то какую бы стратегию j не применил второй игрок, первый получит

$$a_{i_0 j} \geqslant \min_{1 \leqslant j \leqslant n} a_{i_0 j} = \underline{v}_{i_0} = \underline{v},$$

т. е. $\underline{v}$ есть максимальный гарантированный выигрыш первого игрока и первый игрок всегда может обеспечить себе как минимум этот выигрыш.

С другой стороны, если обозначить

$$\overline{v}_j = \max_{1 \leqslant i \leqslant m} a_{ij}$$

и взять j_0 такое, чтобы

$$\overline{v}_{j_0} = \min_{1 \leqslant j \leqslant n} \overline{v}_j = \min_{1 \leqslant j \leqslant n} \max_{1 \leqslant i \leqslant m} a_{ij} = \overline{v},$$

то, выбирая стратегию j_0, второй игрок заплатит первому

$$a_{ij_0} \leqslant \max_{1 \leqslant i \leqslant m} a_{ij_0} = \overline{v}_{j_0} = \overline{v},$$

т. е. второй игрок может проиграть при разумном поведении не больше $\overline{v}$. Значит, первый игрок, предполагая разумность своего противника, может ожидать выигрыш не больший $\overline{v}$.

Давайте посмотрим, как это выглядит на примере матричной игры, которую написал Борис.

Здесь

$$\underline{v}_1 = \min\{2;\ -1;\ -2\} = -2;$$

$$\underline{v}_2 = \min\{0;\ 1;\ 2\} = 0$$

и $$\underline{v} = \max\{-2;\ 0\} = 0,\ i_0 = 2;$$

$$\overline{v}_1 = \max\{2;\ 0\} = 2;$$

$$\overline{v}_2 = \max\{-1;\ 1\} = 1;$$

$$\overline{v}_3 = \max\{-2;\ 2\} = 2$$

и $$\overline{v} = \min\{2;\ 1;\ 2\} = 1,\ j_0 = 2.$$

Таблица 5.4

i \ j	1	2	3
1	3	0	—1
2	—1	1	2
3	0	3	1

Таким образом, первый игрок может гарантировать себе выигрыш не меньше 0, выбирая стратегию $i_0=2$, а второй игрок может гарантировать себе проигрыш не больше 1, выбирая стратегию $j_0=2$.

— Смотрите-ка! — воскликнул Борис, — у нас получилось $\underline{v}<\overline{v}$ и $a_{i_0j_0}=\overline{v}$.

— Что ж, разберемся в этом. То, что $\underline{v}\leqslant\overline{v}$ не удивительно. Это ясно даже по смыслу величин $\underline{v}$ и $\overline{v}$. Но и формально можно получить это неравенство.

Раз $\underline{v}_i\leqslant a_{ij}\leqslant\overline{v}_j$ при любых i и j и $\underline{v}_i$ не зависит от j, а $\overline{v}_j$ — от i, то

$$\underline{v}=\max_{1\leqslant i\leqslant m}\underline{v}_i\leqslant\overline{v}_j \text{ при любом } j \text{ и}$$

$$\underline{v}\leqslant\min_{1\leqslant j\leqslant n}\overline{v}_j=\overline{v}.$$

Величины $\underline{v}$ и $\overline{v}$ называются соответственно нижней и верхней ценами игры. Если они совпадают, то величина $v=\underline{v}=\overline{v}$ называется ценой игры.

Что же касается равенства $a_{i_0j_0}=\overline{v}$, то это получилось случайно. Можно привести пример, когда $a_{i_0j_0}=\underline{v}$ или $\underline{v}<a_{i_0j_0}<\overline{v}$, например, для такой матрицы (табл. 5.4.)

Здесь $\underline{v}=0$, $\overline{v}=2$, $i_0=3$, $j_0=3$, $a_{i_0j_0}=1$.

Вот теперь и возникает главный вопрос — какую стратегию выбирать первому игроку?

— Ну, конечно же, i_0! — воскликнула Галя, — ведь тогда второй выберет j_0 и первый получит $a_{i_0 j_0}$.

— Почему же? — удивленно спросил Стрелкин. — Если быть уверенным, что второй выберет j_0, то, наверно, первый может найти стратегию и получше, чем i_0. Например, в последнем случае (см. табл. 5.4) против стратегии $j_0=3$ второго игрока первому лучше ответить стратегией $i=2$ и получить $a_{23}=2$, чем применить стратегию $i_0=3$, получая $a_{33}=1$.

— Да, но где уверенность в том, что второй применит стратегию $j_0=3$? — спросил Борис.

— Но ведь мы считаем его разумным? Разве не так, Александр Андреевич? — удивилась Галя.

— Конечно, но, в свою очередь, он, считая разумными нас и рассуждая так же, как вы, Галя, что мы применим стратегию $i_0=3$, выберет стратегию $j=1$ и даст нам только $a_{31}=0$.

— Вот так-так! — воскликнул Стрелкин. — Выходит, что предполагая разумность друг друга, первый игрок выберет стратегию $i=2$, чтобы получить $a_{23}=2$, второй игрок выберет $j=1$, чтобы отдать $a_{31}=0$, а в результате первый получит $a_{21}=-1$, т. е. меньше, чем свой наилучший гарантированный результат $\underline{v}=1$. Что ж это за «разумность» такая?

— Видите, — усмехнулся Александр Андреевич. — Вы установили, что одно стремление максимизировать свой выигрыш как принцип оптимальности явно недостаточно. Рассмотрим еще раз пару стратегий (i_0, j_0).

Мы видим, что при фиксированном j_0 у первого игрока есть стратегия i, лучшая, чем i_0, т. е.

$$a_{i j_0} > a_{i_0 j_0},$$

а при фиксированном i_0 у второго игрока есть стратегия j, лучшая, чем j_0, т. е.

$$a_{i_0 j_0} < a_{i_0 j}.$$

Поэтому стратегии (i_0, j_0) образуют, как говорят, неустойчивую пару. Если бы для любых i и j выполнялись неравенства

$$a_{i j_0} \leqslant a_{i_0 j_0} \ll a_{i_0 j},$$

то ни одному из игроков не было бы выгодно отклоняться от стратегий i_0 и j_0. Такую пару стратегий называют седловой точкой в антагонистической игре. Как мы видели в последнем примере, седловой точки может и не быть вовсе. Но если такая точка (i_0, j_0) есть, то можно показать, что тогда игра имеет цену, т. е. $\underline{v} = \overline{v} = v$ и, более того, $a_{i_0 j_0} = v$.

— Теперь ясно, что если в игре есть седловая точка, то i_0 можно считать оптимальной стратегией первого игрока, а j_0 — второго. Верно? — спросил Борис.

— Верно. Остается все же решить, что делать в случае, когда седловой точки нет.

— И еще один вопрос, — вставил Стрелкин, — что если в игре несколько седловых точек?

— Такие игры действительно существуют. Например, с такой матрицей (табл. 5.5).

Здесь $v = \underline{v} = \overline{v} = 1$ и пары (2, 1), (4, 3) образуют седловые точки. Но дело в том, что пары (2, 3) и (4, 1) тоже образуют седловые точки и, следовательно, каждый игрок может применять любую из своих «оптимальных» стратегий независимо от другого. Это свойство называется взаимозаменяемостью оптимальных стратегий в антагонистических играх.

Если же седловых точек нет, то, как мы видели, задача выбора рационального (а не оптимального!) действия

Таблица 5.5

i \ j	1	2	3	4
1	0	3	—1	3
2	1	2	1	3
3	—2	3	0	3
4	1	4	1	2

становится более трудной. Здесь можно выдвигать другие принципы «разумности» поведения, отличные от простого стремления к увеличению своего выигрыша. Можно, например, стать на позиции осторожного игрока и руководствоваться «принципом гарантированного результата», стараясь обеспечить себе максимальную нижнюю оценку выигрыша, возможного при любых стратегиях противника. В игре с матрицей, изображенной в табл. 5.4, в этом случае рациональным выбором первого игрока будет $i=3$. Но, оказывается, есть еще одна возможность поведения в матричных играх, не имеющих седловых точек. Эта возможность заключается в применении смешанных стратегий.

— Мы уже говорили, что смешанные стратегии — случайное использование чистых. Так?— спросила Галя.

— Да, смешанной стратегией первого игрока можно считать вектор $\bar{p}=(p_1, \ldots, p_m)$, где $p_i \geqslant 0$ и $p_1+\ldots+p_m=1$, и понимать это так: с вероятностью p_i первый игрок применяет стратегию с номером i. Точно так же смешанную стратегию второго игрока можно задать вектором $\bar{q}=(q_1, \ldots, q_n)$, где $q_j \geqslant 0$ и $q_1+\ldots+q_n=1$.

— А как вычислить выигрыш при использовании игроками смешанных стратегий? — спросил Стрелкин.

— Так же как в общем случае, о котором мы говорили раньше. Теперь каждая ситуация (i, j) будет реализовываться случайным образом — событие A_{ij}, состоящее в применении первым игроком стратегии i, а вторым — стратегии j будет случайным и наступит с вероятностью $p_i q_j$, так как игроки действуют независимо. Поэтому выигрыш первого игрока будет величиной случайной, принимающей значение a_{ij} при наступлении события A_{ij}, т. е. с вероятностью $p_i q_j$, а математическое ожидание выигрыша будет соответственно:

$$a(\bar{p}, \bar{q}) = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} p_i q_j.$$

Вот эту величину $a(\bar{p}, \bar{q})$ и будет теперь стараться увеличить первый игрок, выбирая p, и уменьшить второй игрок, выбирая q.

— Так что же, мы пришли к той же задаче, что и раньше, но теперь у нас p и q могут пробегать бесконечные множества, — заметил Борис. — Чем же это лучше?

— Дело в том, что справедлива следующая основная теорема теории матричных игр. Всякая матричная игра обязательно имеет седловую точку в смешанных стратегиях, т. е. существуют такие p_0 и q_0, что

$$a(\bar{p}, \bar{q}_0) \leqslant a(\bar{p}_0, \bar{q}_0) \leqslant a(\bar{p}_0, \bar{q})$$

при любых $\bar{p}$ и $\bar{q}$.

Только остается открытым вопрос о целесообразности применения смешанных стратегий. Во всяком случае, если игра проводится многократно, это имеет смысл.

Перейдем теперь к неантагонистическим играм, т. е. к играм, в которых интересы игроков непротивополож-

ны. Мы по-прежнему будем рассматривать случай двух игроков и конечных множеств стратегий обоих. Такую игру удобно записывать в виде двойной матрицы (такие игры называются биматричными) (табл. 5.6).

Здесь величина a_{ij} есть выигрыш первого, а b_{ij} — второго игроков при применении ими стратегий i и j соответственно.

В биматричных играх тоже можно определить устойчивые пары стратегий, аналогичные седловым точкам в антагонистических играх, с помощью понятия ситуаций равновесия, о которых я уже говорил. Здесь условие равновесия пары (i_0, j_0) будет выглядеть так:

$$a_{ij_0} \leqslant a_{i_0j_0} \text{ для любых } i$$

и

$$b_{i_0J} \leqslant b_{i_0j_0} \text{ для любых } j.$$

Так же как и в антагонистических играх, ситуации равновесия в чистых стратегиях могут не существовать,

Таблица 5.6

i \ j	1	. . .	j	. . .	n
1	(a_{11}, b_{11})	. . .	(a_{1j}, b_{1j})	. . .	(a_{1n}, b_{1n})
.	.		.		.
.	.		.		.
.	.		.		.
i	(a_{i1}, b_{i1})	. . .	(a_{ij}, b_{ij})	. . .	(a_{in}, b_{in})
.	.		.		.
.	.		.		.
.	.		.		.
m	(a_{m1}, b_{m1})	. . .	(a_{mj}, b_{mj})	. . .	(a_{mn}, b_{mn})

Таблица 5.7

i \ j	1	2
1	(5,5)	(0,10)
2	(10,0)	(1,1)

но если перейти к смешанным стратегиям, то всегда существует пара $(\bar{p}_0, \bar{q}_0)$ такая, что

$$a(\bar{p}, \bar{q}_0) \leqslant a(\bar{p}_0, \bar{q}_0) \text{ для любых } \bar{p};$$
$$b(\bar{p}_0, \bar{q}) \leqslant b(\bar{p}_0, \bar{q}_0) \text{ для любых } \bar{q}.$$

Однако в неантагонистических играх точки равновесия уже не обладают свойством взаимозаменяемости и выигрыши игроков могут быть различны в разных точках равновесия. Это, конечно, существенный недостаток. Хотя никто из игроков не будет отклоняться от равновесных стратегий, если уж они выбраны, но для того чтобы их выбрать, им нужно принять совместное решение о том, какая из ситуаций равновесия будет разыгрываться (если такая не одна).

Кроме того, здесь возникает неудобство, связанное с непаретовостью и, следовательно, с возможностью улучшить результаты одновременно для обоих игроков при совместных действиях.

Эти недостатки ситуаций равновесия можно преодолевать различными способами. Для того чтобы с ними познакомиться, нужно заняться этим серьезнее.

Вот, например, биматричная игра с табл. 5.7.

Здесь, как легко видеть, ситуацией равновесия является пара стратегий $i_0 = 2$, $j_0 = 2$, и $a_{i_0 j_0} = b_{i_0 j_0} = 1$,

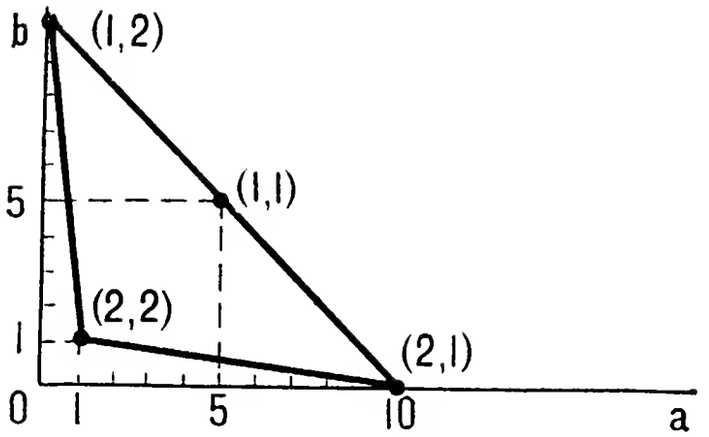


Рис. 5.4

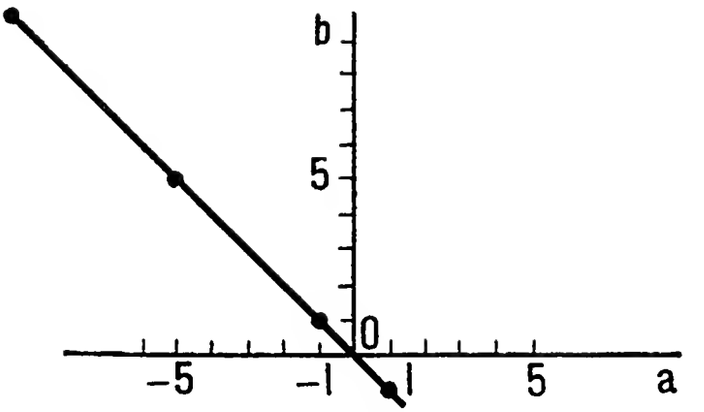


Рис. 5.5

в то время как $a_{11}=b_{11}=5>1$. Пара стратегий $i=1$, $j=1$ дает лучшие результаты обоим игрокам, но не является устойчивой, так как каждому игроку выгодно применить свою вторую стратегию против первой стратегии партнера. Если изобразить на плоскости все возможные исходы (пары выигрышей) в этой игре, то получится следующая картинка (рис. 5.4). На ней ясно видно, что стратегии $i=1$, $j=1$ дают паретовскую точку в этой игре, а ситуация равновесия $i_0=2$, $j_0=2$ далеко не паретовская.

В антагонистической игре все значительно проще. Ведь там $b_{ij}=-a_{ij}$ и соответствующая картинка всегда имеет вид (рис. 5.5). Таким образом, в антагонистической игре любая пара стратегий, и, в частности, седловая, дает паретовское решение.

Итак, в неантагонистической игре возникают две проблемы. Во-первых, поскольку равновесные стратегии не обладают свойством взаимозаменяемости, то надо научиться выбирать из них лучшую. Во-вторых, хотелось бы превращать паретовские точки, хотя бы некоторые, в равновесные.

— Да, — перебил математика Борис, — но ведь с точки зрения разных игроков «лучшими» будут разные ситуации равновесия, так что надо еще определить, какая ситуация лучше.

— Правильно, и здесь выступает один из основных принципов исследования операций. Когда мы начинаем изучать некоторый процесс и строить модель, мы должны описывать его по возможности объективно, учитывая все элементы и их связи, но как только мы переходим к этапу выбора действий, оценивая их эффективность, мы должны твердо установить, с точки зрения какого из элементов модели мы рассматриваем моделируемый процесс, какой критерий выступает для нас основной мерой качества решения. В игровой ситуации это означает — с точки зрения какого игрока мы оцениваем последствия принятых решений.

— Но разве это не означает, — спросил Стрелкин,— что мы как бы ставим этого игрока над другими и нарушаем равноправие, которое было с самого начала.

— Да, это так, но только в этом случае мы сможем давать рекомендации, что делать, именно этому выбранному игроку. Этот принцип фактически означает, что при выработке рекомендаций исследователь операции не может оставаться лицом нейтральным. Все станет на свои места, если считать, что лицо, принимающее решение, и исследователь операций — разные люди. Исследователь операций строит модель, анализирует ее и сообщает лицу, принимающему решение, о результатах исследования, его допустимых действиях в той или иной обстановке и о возможных последствиях. А уж непосредственный выбор осуществляет лицо, принимающее решение в соответствии со своими интересами. Так что «неравноправие» интересов теперь становится естественным — каждого игрока в первую очередь интересует его собственный критерий, который, конечно, может включать в себя и выражение общественной пользы.

— Понятно, — согласился Стрелкин. — Значит мы установили какой-то приоритет среди участников?

Устроили что-то вроде «иерархической системы», о которой сейчас так много говорят.

— Да-да, — усмехнулся Александр Андреевич. — Об иерархических системах стало модным говорить. Но этому есть свое оправдание. Мы всегда подразделяем любую изучаемую систему, любой процесс на главные и второстепенные части, устанавливая тем самым иерархию или приоритет интересов. Но это еще не все. Иерархичность в нашем подходе носит более глубокий характер. Становясь на точку зрения того или иного элемента моделируемой системы, мы всю информацию о системе будем анализировать с его точки зрения и дадим ему право первоочередного выбора действия. В играх это будет означать право первого хода или передачи информации о своем выборе другим игрокам.

— Но разве такое право всегда осуществимо? — удивился Стрелкин.

— Конечно, нет. Но теперь мы рассмотрим именно такие игровые ситуации. Будем их называть играми с иерархической структурой. Если игровая ситуация такого рода включает только двух участников, то выделим из них одного, которого в дальнейшем будем называть первым, и будем искать «наилучшие» с его точки зрения решения, предоставив ему право передавать информацию о своих действиях второму игроку. Давайте построим математическую модель такой игры.

Как и раньше, у нас будут два игрока с множествами стратегии X — первого и Y — второго и функциями выигрышей (критериями) $F(x, y)$ и $G(x, y)$ соответственно. Отличие в том, что первый игрок теперь имеет возможность в той или иной форме передавать информацию о своем выборе второму игроку.

— Зачем же это ему нужно? — удивилась Галя. — Если двое играют, то зачем же одному открывать другому свои намерения?

— Но ведь вы, Галя, обсуждаете с подругами свои планы, хотя ваши интересы не всегда совпадают, верно?

— Так ведь то — с подругами и потом именно обсуждаем: я им говорю о своих планах, а они мне о своих.

— Здесь, конечно, не совсем то же самое. Но в качестве «ответного сообщения» второго игрока может выступить предположение о его разумности и «не вредности», т. е. предположение, что второй игрок имеет единственной своей целью — увеличить свой выигрыш. Чтобы это стало ясней, давайте рассмотрим конкретные способы обмена информацией между игроками.

— Что значит — способы? — спросил Стрелкин. — Один способ и есть: сообщил свое x или не сообщил.

— Нет, это было бы слишком просто, — возразил математик. — Информацию можно передавать в различных видах и вы сами, наверно, с этим встречались. Самый простой и часто встречающийся пример такого рода должен быть связан с распределением какого-нибудь ресурса, находящегося во власти верхнего уровня, причем само распределение может зависеть от предполагаемых действий нижнего уровня, и информация, передаваемая нижнему уровню, может отражать различные формы такой зависимости.

— Я, кажется, понял. И даже могу указать практический пример. Вся железнодорожная сеть в нашей стране делится на так называемые «дороги». Высшим органом каждой «дороги» является управление, обладающее некоторой автономией действия и решающее основные вопросы планирования и руководства работой этой части железнодорожной сети. Отделение дороги, а их в подчинении у одного управления может быть несколько, решают задачи организации перевозок на своем подразделении.

Если интересы управления дороги объединяют многие характеристики, то отделения заинтересованы в получении максимальной прибыли от перевозок грузов на своих участках. Для перевозки нужны порожние вагоны. А их распределением ведает управление. Вот тут-то и возникает конфликт.

— Но не антагонистический, — подчеркнул Александр Андреевич. — Ведь управление дороги тоже заинтересовано в прибылях.

— Да, но оно также заинтересовано в уменьшении времени простоя порожних вагонов, — продолжил Стрелкин. — И вот между отделом управления, ведающим распределением порожних вагонов, и каким-то отделением возможны такие формы отношений.

Управление может выделить определенное количество вагонов для отделения без каких-либо обсуждений, а отделение уже будет решать свои задачи при этих выделенных ресурсах. Другая возможность состоит в том, что управление выделит вагоны в зависимости от тех действий, которые предпринимает отделение дороги (например, от выполнения плана в прошлом). И, наконец, может быть так, что управление дороги определяет количество выделенных отделению порожних вагонов в зависимости от плана их использования.

— Прекрасно! — воскликнул Александр Андреевич. — Давайте все это опишем математически.

Итак, первая форма передачи информации от первого игрока (в вашем примере — управление дороги) состоит в непосредственном сообщении своей стратегии x (количество порожних вагонов, выделяемых отделению). Игру с такой передачей информации обозначим Γ_1. Вторая форма состоит в том, что первый игрок сообщает второму x как функцию от y (действия отделения), т. е. стратегии первого игрока теперь задаются в виде $x=f(y)$. Такую игру будем обозначать Γ_2.

Наконец, третья форма может быть выражена так. Если первый игрок сообщит второму x (управление выделит отделению x вагонов), то отделение распорядится ими с помощью действий y, зависящих от x, т. е. стратегии второго игрока задаются в виде функций $y=h(x)$. Поэтому стратегии первого игрока будут «функциями от функций» $x=\varphi(h)$. Эту игру будем обозначать Γ_3. Посмотрим, как же осуществляются указанные стратегии в действительности.

Игра Γ_1

Узнав от первого игрока его стратегию x, второй будет выбирать y так, чтобы

$$G(x, y)=\max_{y'\in Y} G(x, y'), \tag{5.3}$$

т. е. выберет наилучший со своей точки зрения ответ.

— Но ведь не всегда такое y существует, — вставил Борис. — Максимум ведь может не достигаться.

— Верно, — ответил Александр Андреевич. — Но мы сейчас не будем останавливаться на таких тонкостях. Предположим, что такое y существует.

— А если ответов много? — не успокоился Борис.

— Это уже более важно. Давайте обозначим через $Y(x)$ множество всех тех y, которые при данном x для второго игрока наилучшие, т. е. такие, для которых выполнено равенство (5.3).

Игра Γ_2

Первый игрок сообщил второму $f(y)$, т. е. если второй игрок выберет y, то первый применит стратегию $x=f(y)$ и второй получит $G(x, y)=G(f(y), y)$.

Обозначим через $Y(f)$ множество таких y, для которых $G(f(y), y) = \max\limits_{y' \in Y} G(f(y'), y')$, т. е. множество лучших ответов второго игрока[a]. Опять же будем считать, что для любой функции f такие y существуют и множество таких y обозначим через $Y(f)$.

Игра Γ_3

Первый игрок сообщает второму «функцию от функции» $\varphi(h)$. Если второй выберет функцию $h(x)$, то первый применит стратегию $x = \varphi(h)$ и второй реализует свой план $y = h(x)$. Таким образом, второй игрок получит $G(x, y) = G(\varphi(h), h(x))$, где $x = \varphi(h)$, и «наилучшей» его стратегией будет такая функция h, для которой

$$G(\varphi(h), h(x)) = \max_{h'} G(\varphi(h'), h'(x)).$$

Множество таких h обозначим через $H(\varphi)$.

Теперь вернемся к первому игроку. Какую же стратегию ему выбирать? Естественно, что предположения о разумности второго игрока в нашем случае означают, что второй игрок будет выбирать свою стратегию соответственно из множеств $Y(x)$, $Y(f)$ и $H(\varphi)$. Но все стратегии каждого из этих множеств в соответствующей игре одинаковы с точки зрения выигрыша второго игрока. Поэтому у первого нет никаких оснований предполагать, что второй предпочтет какую-то определенную из них. Значит, для первого игрока неопределенным фактором является выбор вторым игроком стратегии внутри его множества, соответствующего типу игры.

Выбрав в игре Γ_1 стратегию x, первый игрок может гарантировать себе только $\min\limits_{y \in Y(x)} F(x, y)$.

Следовательно, его максимальный гарантированный результат равен:

$$v_1 = \max_{x \in X} \min_{y \in Y(x)} F(x, y)$$

и его оптимальной стратегией является то x, на котором этот максимум достигается.

Аналогично максимальный гарантированный результат в играх Γ_2 и Γ_3 определяется формулами:

$$v_2 = \max_{f} \min_{y \in Y(f)} F(f(y), y)$$

и

$$v_3 = \max_{\varphi} \min_{h \in H(\varphi)} F(\varphi(h), h(x)), \text{ где } x = \varphi(h).$$

— Все стало вдруг так сложно, — сказала Галя.

— Постойте-ка! — воскликнул Борис. — Я, кажется, понял. Ведь если первый игрок сообщит второму x_0, которое отвечает ситуации равновесия, лучшей именно для него, первого, то второй игрок вынужден будет тоже согласиться на нее и использовать ту стратегию y_0, которая вместе с x_0 эту ситуацию равновесия и составляет. Ведь иначе

$$G(x_0, y) \leqslant G(x_0, y_0)$$

для любого y по определению ситуации равновесия и второму невыгодно от нее отклоняться.

— Молодец, Боря, — похвалил Александр Андреевич. — Мы действительно получили, что уже в игре Γ_1 первый игрок может обеспечить себе самую «лучшую» с его точки зрения ситуацию равновесия. Но дело в том, что всегда выполняется неравенство

$$v_1 \leqslant v_3 \leqslant v_2. \qquad (5.4)$$

Поэтому в играх Γ_2 и Γ_3 первый игрок может гарантировать себе еще лучшие результаты.

— Подождите-ка, — вставил Стрелкин. — Вы не ошиблись, что $v_3 \leqslant v_2$, а не наоборот? Ведь это означает в моем примере, что управлению дороги выгоднее не узнавать предполагаемый план использования выделяемых вагонов отделением, а указывать сразу свою реакцию на все возможные действия отделения.

— Все правильно. Здесь нужно учесть, что в игре Γ_3 у первого игрока вроде бы больше возможностей, чем в Γ_2, но зато и у второго игрока в Γ_3 больше возможностей. В результате же оказывается, что $v_3 \leqslant v_2$.

— Ну хорошо. С ситуациями равновесия все ясно. А как быть с паретовостью? — спросил Борис.

— Этот вопрос сложнее. Для того чтобы сделать устойчивыми паретовские решения, нужно переходить к более сложным моделям, и, в частности, учитывать повторения игровых ситуаций. Мы сейчас не будем останавливаться на этом.

— К тому же время для наших бесед, кажется, исчерпано, — сказал Александр Андреевич, взглянув в окно. — Мы затронули лишь часть проблем современной прикладной математики. Если они вас заинтересовали, постарайтесь сами глубже разобраться в них. Желаю вам успеха...

Поезд медленно подходил к вокзалу большого сибирского города.

СПИСОК ЛИТЕРАТУРЫ

Борель Э. Вероятность и достоверность. М., «Наука», 1964, 119 с.

Бусленко Н. П., Шрейдер Ю. А. Метод статистических испытаний (Монте-Карло) и его реализация на цифровых вычислительных машинах. М., «Физматгиз», 1961, 226 с.

Вагнер Г. Основы исследования операций, т. 1. М., «Мир», 1972, 335 с.

Вентцель Е. С. Теория вероятностей. М., «Наука», 1969, 576 с.

Вентцель Е. С. Теория вероятностей (Первые шаги). М, «Знание», 1977. 64 с.

Вентцель Е. С. Исследование операций. М., «Знание», 1976, 64 с.

Гнеденко Б. В., Хинчин А. Я. Элементарное введение в теорию вероятностей. М., «Наука», 1976. 168 с.

Гнеденко В. В., Коваленко И. Н. Введение в теорию массового обслуживания. М., «Наука», 1966, 431 с.

Кокс Д., Смит У. Теория очередей. М., «Мир», 1966, 218 с.

Кофман А., Крюон Р. Массовое обслуживание теория и приложения. М., «Мир», 1965, 302 с.

Льюис Р. Д., Райфа Х. Игры и решения. Введение и критический обзор. М., «ИЛ», 1961, 642 с.

Мак-Кинси Дж. Введение в теорию игр. М., «Физматгиз», 1960, 420 с.

Мостеллер Ф, Рурке Р, Томас Дж. Вероятность. М., «Мир», 1969, 431 с.

Овчаров Л. А. Прикладные задачи теории массового обслуживания. М., «Машиностроение», 1969, 324 с.

Прабху Н. Методы теории массового обслуживания и управления запасами. М., «Машиностроение», 1969, 356 с.

Реньи А. А. Письма о вероятности. М., «Мир», 1970, 93 с.

Риордан Д. Вероятностные системы обслуживания. М., «Связь», 1966, 184 с.

Розенберг В. Я., Прохоров А. И. Что такое теория массового обслуживания. М., «Советское радио», 1965, 256 с.

Саати Т. Л. Элементы теории массового обслуживания и ее приложения. М., «Советское радио», 1971, 520 с.

Соболь И. М. Метод Монте-Карло. М., «Наука», 1968, 64 с.

Хинчин А. Я. Работы по математической теории массового обслуживания. М., «Физматгиз», 1963, 235 с.

ОГЛАВЛЕНИЕ

Приглашение к беседам 4

БЕСЕДА ПЕРВАЯ

Закономерность случайностей. (Теория вероятностей) 10

БЕСЕДА ВТОРАЯ

Станем в очередь. (Теория массового обслуживания) 82

БЕСЕДА ТРЕТЬЯ

Чем сложнее, тем проще. (Динамика средних). 166

БЕСЕДА ЧЕТВЕРТАЯ

Бросим жребий. (Метод Монте-Карло) 180

БЕСЕДА ПЯТАЯ

Какие решения принимать? (Математическое программирование и теория игр) 192

Список литературы 239

Герман Александрович Платонов
Михаил Александрович Файнберг
Михаил Самуилович Штильман

ПОЕЗДА, ПАССАЖИРЫ... И МАТЕМАТИКА

Рецензенты: *В. А. Кудрявцев, В. Н. Воскресенский*
Редактор *И. Б. Сотников*
Обложка художника *Ю. А. Черепанова*
Технический редактор *О. Н. Крайнова*
Корректор *В. Н. Яговкина*
ИБ № 951

Сдано в набор 22/II 1977 г. Подписано к печати 2/IX 1977 г. Формат 70×108 1/32. Бум. тип. № 1. Печ. л. 7,5 (усл. 10,5). Уч.-изд. л. 9,26. Тираж 15 000. Т-15664. Изд. № 1-5-3/4 № 5442. Зак. тип. 1267. Цена 60 коп.
Изд-во «ТРАНСПОРТ», Москва, Басманный туп., 6а

г. Куйбышев, пр. К. Маркса, 201.
Тип. изд-ва «Волжская коммуна».

60 коп.

Поезда, пассажиры... и математика.
Казалось бы, что между ними общего?
Нам, авторам книги, железнодорожникам
и математикам, на своем опыте пришлось
убедиться в правомерности
и плодотворности союза математики
с организацией транспортного процесса.
Поэтому мы попытались привлечь внимание
широкого круга железнодорожников
к возможностям математических
теорий и методов. Мы сознательно
обращались только к тем из них,
которые уже проверены практикой
работы железных дорог. Нам хотелось,
чтобы об этих методах и теориях узнали
подробнее, осмыслили их суть.
Трудно рассказывать о любой науке
популярно, еще труднее популяризировать
математику. Не надеясь на свои силы,
мы решили привлечь к нашему рассказу
помощников. Так появились пассажиры.
Чтобы времени для бесед было достаточно,
поезд отправили в один из далеких городов.
Вот почему в нашей книге встретились
и ПОЕЗДА, и ПАССАЖИРЫ, и МАТЕМАТИКА!